농학박사 학위논문

근적외선 분광분석법과 인공신경망에 의한 목재의 수종 구분

Classification of Wood Species using Near-infrared Spectroscopy and Artificial Neural Networks

2018년 11월

서울대학교 대학원

환경재료과학전공

양 상 윤

근적외선 분광분석법과 인공신경망에 의한 목재의 수종 구분

Classification of Wood Species using Near-infrared Spectroscopy and Artificial Neural Networks

지도교수 여 환 명

이 논문을 농학박사 학위논문으로 제출함

2018년 11월

서울대학교 대학원

환경재료과학전공

양 상 윤

양상윤의 농학박사 학위논문을 인준함

2019년 1월

위 원 장 오 정 권 (인)

부위원장 여 환 명 (인)

위 원 최 인 규 (인)

위 원 심 국 보 (인)

위 원 권 오 경 (인)

초록

근적외선 분광분석법과 인공신경망에 의한 목재의 수종 구분

Classification of Wood Species using Near-infrared Spectroscopy and Artificial Neural Networks

서울대학교 대학원
산림과학부 환경재료과학전공
양상윤

현미경을 이용한 전통적인 조직학적 수종 식별 방법은 목재의 재색, 표면 특징, 해부학적 구조 등을 명확하게 구분함으로써 수종 식별을 실시한다. 따라서, 해부학적 수종 구분을 위해서는 충분히 숙련된 목재 해부학 전문가가 필요하다. DNA 분석에 의한 수종 식별 방법은 목재의 세포 조직으로부터 세포핵 추출이 어려워 분석이 불가능한 경우가 존재하며, 분석에 많은 비용이 소요된다. 본 연구에서는 목재의 수종을 간편하고 신속하게 구분하기 위하여 근적외선 분광분석법과 인공신경망을 적용한 수종 구분 방법을 개발하였다. 국내 제재목 생산업의 침엽수 소비량 중 대다수를 차지하는 낙엽송, 소나무, 잣나무, 삼나무 및 편백의 수종을 구분하기 위해 제재목에서 근적외선 스펙트럼을 측정하여 다양한 알고리즘에 의해 수종 구분을 실시하였다.

근적외선 스펙트럼을 이용한 주성분 분석(Principal component analysis)과 이에 기반한 soft independent modelling of class analogy(SIMCA)를 실시하였다. 근적외선 스펙트럼의 3가지 수학적 전처리 조건(Raw, standard normal variate (SNV), Savitzky-Golay

2^{nd} derivative (SG 2^{nd}))에 따라 전체 스펙트럼을 이용한 주성분 분석을 실시한 결과, PC1과 PC2에 속하는 score의 중첩에 의해 주성분 분석을 이용한 수종 구분은 불가능하였다. 각 수종 집단별 주성분 분석에 의한 SIMCA 분류를 실시한 결과, 수학적 전처리 조건에 따른 수종 구분 신뢰도에 차이가 있었다. SG 2^{nd} 전처리를 실시한 스펙트럼을 이용한 SIMCA 분류가 가장 높은 신뢰도를 나타내었으며, 정확도는 73.00%, 최소 정밀도는 98.54%, 최소 재현율은 67.50%로 평가되었다.

부분 최소 자승 판별 분석은 집단에 따라 1 또는 0이 되는 모조 종속 변수(dummy dependent variable)를 갖도록 하는 다중 선형 회귀 모델을 개발함으로써 분류를 실시하는 방법이다. 교차 검정 예측치 0.5를 기준으로 동일한 3가지 수학적 전처리 조건에 따른 다중 수종 판별을 실시한 결과, 수학적 전처리 조건에 따른 수종 구분 신뢰도에 차이가 나타났으며, SNV 전처리는 부분 최소 자승 판별에 부정적인 영향을 미치는 것으로 판단되었다. SG 2^{nd} 전처리를 실시한 스펙트럼을 이용한 부분 최소 자승 판별 모델의 정확도는 74.9%, 최소 정밀도는 100%, 최소 재현율은 69%로 평가되어 가장 개선된 수종 구분 성능을 나타내었다.

각 수종별 부분 최소 자승 판별 분석에 의한 교차 검정 예측치의 분포를 정규분포로 변환하여 확률분포에 의한 수종 구분을 실시한 결과, 원 스펙트럼과 SNV 전처리를 실시한 스펙트럼을 이용한 판별 분석 모델은 성능 개선이 미미한 것으로 나타났으나, SG 2^{nd} 전처리를 실시한 판별 분석 모델의 경우 성능이 대폭 개선되었다. 이 때의 정확도는 95.18%, 최소 정밀도는 99.19%, 최소 재현율은 91.5%로 평가되었다. SG 2차 미분을 실시한 판별 모델의 정확도 향상에 영향한 근적외선 파장대역을 탐색하기 위해 variable importance in projection (VIP) score를 분석한 결과, Cellulose의 흡광영역인 1632 nm, lignin의 흡광영역인 1698 nm, lignin 및 hemicellulose의 흡광영역인 1720 nm 등과, 목재의 주요 성분에 의한 흡광 영역으로 밝혀지지 않은 1895 nm 및 2304 nm 인근이 SG

2차 미분에 의해 강조되면서 수종 구분 성능 개선에 기여한 것으로 평가되었다.

입력 데이터로부터 예측 또는 분류를 수행하는 최적의 가중치를 탐색하는 인공신경망과 최적의 필터를 탐색하는 1차원 합성곱 신경망을 이용하여 근적외선 스펙트럼을 이용한 목재 수종 구분을 실시하였다.

1721-64-5 (입력층-은닉층-출력층) 노드를 갖는 인공신경망을 설계하여 검정세트를 이용한 목재의 수종 구분을 실시한 결과, 3가지 수학적 전처리 조건에서 모두 확률에 기반한 부분 최소 자승 판별 분석과 유사하거나 더 개선된 성능으로 수종 구분이 가능하였다. 특히 SG 2^{nd} 전처리를 실시한 스펙트럼을 이용한 판별 모델의 정확도, 정밀도, 재현율은 모두 100%로 평가되었다. 인공신경망을 이용한 수종 구분 시 수학적 전처리에 따라 분류성능이 개선되는 것을 확인할 수 있었으며, 특히 SG 2^{nd} 전처리를 실시하는 경우 국산 침엽수 5수종의 제재목에서 측정한 근적외선 스펙트럼을 이용하여 충분한 수종 구분이 가능할 것으로 판단되었다.

근적외선 분광분석법에서 사용되는 수학적 전처리에 대한 의존도를 낮추기 위하여 신경망이 직접 수종 구분에 필요한 최적의 전처리 방법을 찾아나가는 1차원 합성곱 신경망을 활용한 수종 구분을 실시하였다. 본 연구에서는 각기 다른 필터 크기와 채널을 갖는 1차원 합성곱층을 4개 배치하여 스펙트럼의 필터링, 분리, 합성에 의한 전처리를 실시하도록 설계하였다. 1차원 합성곱 신경망을 이용하여 검정세트에 의한 수종 구분 정확도를 평가한 결과, 원 스펙트럼을 이용하여 학습한 1차원 합성곱 신경망의 수종 구분 정확도는 99.9%, 최소 정밀도 및 재현율은 99.5%로 나타났다. SNV를 실시한 스펙트럼을 이용하여 학습을 실시한 1차원 합성곱 신경망 또한 동일한 신뢰도로 평가되었다. SG 2^{nd} 전처리를 실시한 스펙트럼을 이용하여 학습을 실시한 1차원 합성곱 신경망은 검정세트의 정확도, 정밀도, 재현율이 모두 100%로 나타났다. 최종적으로, 본 연구에서 적용한 근적외선 스펙트

럼을 이용한 수종 구분 방법 중 신경망 이론, 특히 1차원 합성곱 신경망을 활용하는 경우 가장 정확한 수종 구분이 가능한 것으로 판단된다.

주요어 : 근적외선 분광분석법, 수종 구분, 주성분 분석, 부분 최소 자승 판별 분석, 인공신경망, 1차원 합성곱 신경망

학 번 : 2013-30337

목 차

제 1 장

제 2 장

표 목 차

그 림 목 차

제 1장

서론 및
이론적 배경

1. 서 론

1. 연구 배경 및 목적

목재는 수종별로 다양한 성능을 나타내기 때문에 수종에 따라 다른 가격에 거래된다. 목재는 주로 사용 목적에 맞게 가공되어 시중에 유통되므로, 일반인은 물론 목재를 다루는 전문가 또한 육안으로 정확한 수종 구분이 어렵다. 문제는 목재 수종 구분의 어려움을 상업적으로 악용하는 목재 유통업체들이 존재한다는 점이다. 국내 목조 주택 및 조경 시장이 확장됨에 따라, 목재를 이용한 데크, 루바, 사이딩 등 목재제품이 시장으로 공급되고 있다. 그러나, 소비자에게 수종을 명확하게 알리지 않거나 다르게 표기한 채 유통하는 문제가 발생하고 있다. 목재 유통업체의 고의성 유무를 떠나, 잘못된 수종 표기는 목제품 성능에 대한 신뢰성 하락을 야기하여 목재 산업의 발전을 위축시킬 가능성이 높다.

상기와 같은 문제를 미연에 방지하고 건전한 목재산업의 발전을 촉진하기 위하여, 2017년 재개정되어 시행중인 목재의 지속가능한 이용에 관한 법률(법률 제14657호)은 목재제품의 규격과 품질 검사 결과를 의무적으로 표시하도록 규정하고 있으며, 규격 및 품질의 기준은 동법의 행정규칙(대통령령 제 28311호)에 고시되어 있다. 이 중 제재목은 Fig. 1-1과 같이 용도, 등급, 수종, 원산지, 치수, 함수율 및 생산자(또는 수입자)가 표시된다. 이 때, 생산 및 유통업체에서 작성하기 어려운 표시 항목은 함수율과 수종이다. 함수율의 경우 고시의 함수율 측정 방법에 부합하도록 시험편을 채취하여 표시 할 수 있으나, 수종 표시는 목재제품 품질표시제도에 의한 수종검사에 의해 실시된다. 이는 제재 업체와 수입 업체 모두에게 해당된다. 따라서 사용방법이 간편하면서 현장에서 즉시 확인 가능한 수종 구분 기술이 필요하다.

목재의 수종을 식별하기 위해 목재의 해부학적 세포 조직 비교 분

석 또는 DNA 분석 등이 이용되고 있다. 그러나 두 방법 모두 재료를 절단해야하고, 분석에 오랜 시간이 소요된다. 현미경을 이용한 전통적인 조직학적 수종 식별 방법은 각기 다른 목재의 재색, 표면 특징, 해부학적 구조 등을 명확하게 구분할 수 있어야 수종 식별이 가능하다. 따라서, 해부학적 수종 구분에는 충분히 숙련된 목재 해부학 전문가가 필요하다. DNA 분석의 경우 목재의 세포 조직으로부터 세포핵 추출이 어려워 분석이 불가능한 경우도 존재하며, 분석에 많은 비용이 소요된다는 단점이 있다.

표시 사항	품명 - 등급 - 수종 - (원산지) 치수 - 함수율 - 생산(수입)자
예1)	수장용 판재 - 2등급 - 오동나무 - 한국 30㎜×150㎜×2.4m - KD12 - ○○제재소
예2)	보구조재 - E11 - 소나무 180㎜×300㎜×4.0m - KD15(주) - ××상사 (중국)

Figure 1-1. Standard for lumber quality indication

목재는 수종과 생장환경에 따라 물리적 · 화학적 성질이 다르게 나타난다. 그러나 그 물성 평가에는 시간과 비용이 많이 소요되므로 이를 간접적으로 정확하면서 신속하게 분석할 수 있는 기술들이 연구되어왔다. 이 중, 근적외선 분광분석법은 파장범위 780 ~ 2500 nm에서 물체의 화학적 구성요소 차이에 따라 발생하는 흡광도 변화를 스

펙트럼으로서 분석하는 기술로, 목재와 같은 유기물의 특성을 분석하는데 유효한 분석 방법이다. 근적외선 스펙트럼 데이터를 이용하여 다변량 통계분석법을 적용함으로써 근적외선 분광분석법의 활용 범위가 확대되어 농업, 식품 및 사료뿐 아니라 현재는 화학, 의학, 약학, 생화학, 석유화학, 고분자 및 섬유분야까지 널리 보급되어 정량·정성분석에 활발히 적용되고 있다(Blanco, 2002; Pasquini, 2003; Luypaert, 2007;. Porep et al, 2015). 목재 분야에서도 다양한 형태(제재목/칩/목분)에서 측정한 근적외선 스펙트럼을 이용한 물성 및 품질 분석에 관한 연구들이 활발히 수행중에 있다. 이 중, 근적외선 스펙트럼을 이용한 목재의 수종 구분 가능성을 제시하는 연구들이 보고되고 있다(Pastore et al., 2011, Braga et al., 2011, Bergo et al., 2016, Nisgoski et al., 2017, Lazarescu et al. 2017). 그러나 많은 연구들이 시험편의 분말화 또는 칩형태의 물리적 전처리를 실시한 시험편으로부터 측정한 근적외선 스펙트럼을 이용하여 수종 구분을 실시하거나, 소량의 스펙트럼 데이터를 이용하여 구분을 실시하였기 때문에 그 활용범위가 제한적이었다.

본 연구에서는 생물재료의 비파괴적 평가에 효과적으로 활용되고 있는 근적외선 분광분석법을 목재 수종 구분에 도입하고자 한다. 근적외선 대역은 시료의 전처리와 같은 과정 없이 비파괴적으로 스펙트럼을 측정할 수 있으며, 측정이 간편하여 편리하게 활용할 수 있다. 또한 측정 시간이 짧아(1분 이내) 대량의 시험편에 대한 측정 및 평가 가능하다. 이에 국내에서 유통량이 많은 국산 주요 침엽수종(낙엽송, 소나무, 잣나무, 삼나무, 편백)을 구분하기 위해 국내 각지에서 수집한 제재목으로부터 근적외선 스펙트럼을 획득하여 근적외선 분광분석 분야에서 전통적으로 활용되어 온 다변량 통계분석 방법인 주성분 분석, soft independent modelling of class analogy(SIMCA), 부분 최소 자승 판별 분석과 최근 딥러닝 등으로 활발히 연구가 수행되고 있는 인공신경망 및 합성곱 신경망을 이용한 국산 침엽수 제재목의 수종 구분을 실시하고자 한다.

2. 연구 목표

본 연구의 목표를 요약하면 다음과 같다.

- 국내 각지에서 주요 침엽수종(낙엽송, 삼나무, 잣나무, 소나무, 편백)을 수집하여 근적외선 스펙트럼을 획득한다.

- 주성분 분석과 SIMCA를 이용한 수종 구분 신뢰도를 평가한다.

- 부분 최소 자승법을 이용한 판별 분석에 의한 수종 구분 신뢰도를 평가한다.

- 인공신경망 및 1차원 합성곱 신경망을 이용한 수종 구분을 실시하고 신뢰도를 평가한다.

- 각 수종 구분 알고리즘간의 신뢰도를 평가하여 근적외선 스펙트럼을 이용한 최적 수종 구분 시스템을 선정한다.

3. 이론적 배경

3.1 근적외선 분광분석법

William Hershel에 의해 적외선(infrared) 영역이 발견된 이후 1890년대 플랑크(N. Plank)의 흑체이론에 의해 적외선을 이용한 온도 측정이 가능해지면서 적외선 영역에 관한 연구 개발이 활발하게 이루어지기 시작하였다. 적외선은 파장에 따라 세 가지 영역으로 분리하여 명명하고 있으며, 그 경계는 기관에 따라 정의에 차이가 있다. 국제표준화기구에서 정의한 ISO 20473-2007에 따르면, 적외선 영역은 Table 1-1과 같이 근적외선(near-infrared), 중적외선(mid-infrared), 원적외선(far-infrared)으로 나눌 수 있다(Fig. I-2).

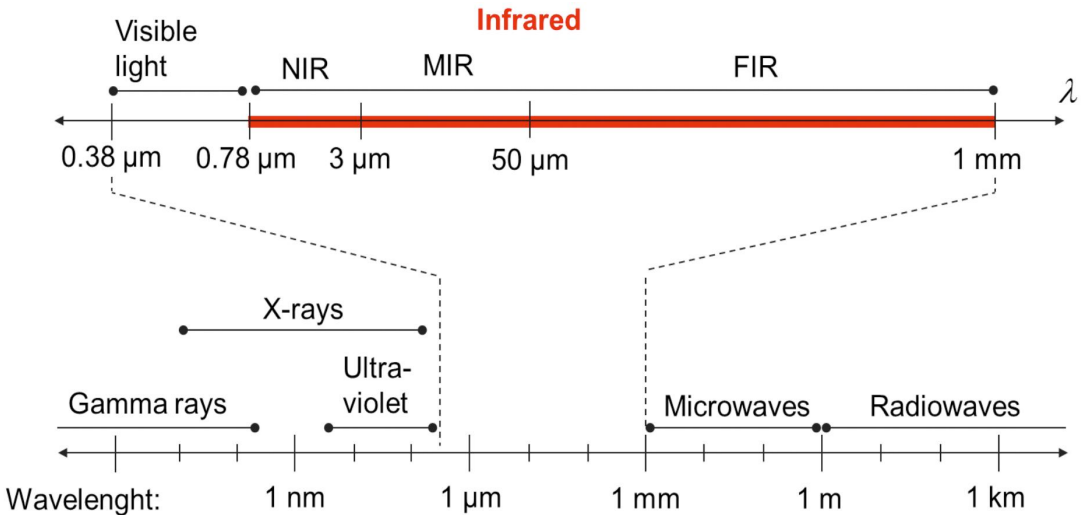


Figure 1-2. Spectral range of electromagnetic waves.

(출처:https://www.petnology.com/competence-online/news/preform-heating-methods-in-the-two-stage-stretch-blow-moulding-process/detail.html)

Table 1-1. Spectral bands for optics and photonics. (ISO 20473-2007)

	Wavelength(nm)	Wavenumber(cm^{-1})
Near-Infrared	780 ~ 3,000	13,000 ~ 7,000
Mid-Infrared	3,000 ~ 50,000	3,300 ~ 200
Far-Infrared	50,000 ~ 1,000,000	200 ~ 10

물체에 적외선을 조사하면 물체를 구성하는 분자들은 에너지가 더 높은 상태로 들뜨게(exited) 된다. 적외선은 공유결합 분자에서 결합의 신축(stretching) 및 굽힘(bending) 진동을 유발하는 정도의 에너지 준위를 갖는다. 분자 내 작용기들은 적외선의 진동수가 분자의 고유 진동수와 일치할 때 적외선을 흡수하고, 흡수된 에너지로 인하여 분자 내 결합의 진동 운동의 진폭(amplitude)은 증가한다. 근적외선 분광분석법에서 활용되고 있는 파장 영역은 대개 780 ~ 2,500 nm 대역으로서, 주로 중적외선에서 유래되는 -CH, -OH, -NH 작용기의 배음대(overtone band)와 결합대(combination band)에 의한 에너지 흡수가 주를 이룬다(Pasquini, 2003). 이를 이용하여 근적외선 스펙트럼을 분석함으로써 다양한 유기체의 물성 분석을 실시할 수 있다.

근적외선 대역은 여러 작용기의 결합대와 배음대의 중첩도가 높고, 광흡수대가 넓기 때문에 중적외선에 비해 작용기의 흡수지점을 특정하기 어려워 단일 파장을 이용한 정량분석 또는 정성분석에 어려움이 있다. 반면, 중적외선 영역에 비해 광원을 확보하기 쉽고, 시료에 대한 전처리가 필요하지 않으며, 장치의 구성이 간편하다는 장점이 있다. 근적외선의 도입 초기단계에서는 위와 같은 단점에 의해 근적외선의 활용이 제한적이었으나, 1950년대 미국 농무성의 Karl Norris가 처음으로 농산물의 성분 분석에 근적외선을 활용한 연구 결과를 보고하면서, 근적외선 분광분석기가 발전하기 시작하였다. 1980년대에 이르러 다변량 통계분석에 의한 화학계량학(chemometrics)이 도입되면서 근적외선 분광분석법을 이용한 정량 및 정성분석이 정밀한 수준으로 가능해졌다.

근적외선 스펙트럼의 측정은 크게 투과(transmission)에 의한 흡광도(absorbance) 측정과 확산반사(diffuse reflection)에 의한 흡광도 측정 두

가지 방법으로 실시된다. 일반적으로 산란이 거의 발생하지 않는 액체와 같은 물질은 큐벳에 적정량을 주입하여 기준광의 반사가 발생하지 않도록 수직으로 광원을 입사시킴으로써, 파장별 입사광량 대비 투과광량의 차이를 이용하여 흡광도 스펙트럼을 측정한다. 고체의 경우에도 분말화가 가능한 경우 KBr에 의해 펠렛화하여 투과에 의한 흡광도 측정이 가능하다.

확산반사는 입사된 광이 표면에서 반사되는 직반사(specular reflection)와 물체 내부로 투과된 광이 산란에 의해 입사방향으로 다시 돌아오는 후방 산란(back scattering)현상을 통칭하는 것으로, 확산반사에 의한 근적외선 측정방법에서는 주로 후방 산란광만을 이용하여 스펙트럼을 분석한다(Wang *et al.*, 1995). 근적외선 대역은 광 투과성이 낮아 확산반사에 의해 근적외선 스펙트럼을 측정하는 경우 시료의 두께에 큰 영향을 받지 않는다. 또한 고체 시료를 직접 측정하는 경우, 물리/화학적 전처리를 필요로 하지 않기 때문에 비파괴적으로 스펙트럼을 측정할 수 있다. 또한 여타의 분광분석기에 비해 신속한 스펙트럼 측정가능하다는 장점이 있다.

목재와 같은 고체의 경우 근적외선의 투과가 거의 발생하지 않고, 산란과 확산반사가 크게 발생하기 때문에 대부분의 경우 확산반사에 의한 스펙트럼 측정을 실시한다. 목재 또한 펠렛화 또는 마이크로톰을 이용하여 박편을 획득하여 투과도 측정이 가능하지만, 시험편을 통제해야하는 제한적인 조건에서만 투과도 측정이 실시된다(Tsuchikawa et al., 1998; Tsuchikawa and Sisler, 2003; Yeh et al., 2004; Yeh et al., 2005). 확산반사를 이용한 고체 시료의 근적외선 스펙트럼 측정은 내부 구조에 따라 간섭이 발생하기 때문에 액체 시료에 비해 재현성이 떨어진다. 목재의 경우 직교 이방성을 가지고 있으며 다공성의 세포 구조를 가진 비균질한 고체이므로 방사 또는 접선단면과 횡단면에서의 확산반사 양상이 다르게 나타난다(Kienle et al., 2008).

따라서 근적외선 스펙트럼을 이용한 분석의 신뢰도를 높이기 위해서는 스펙트럼 측정의 재현성을 확보하기 위한 노력이 반드시 수반되어야 한

다. 제재목으로부터 근적외선 스펙트럼을 측정하고자 할 때는 한정된 반사각에서 스펙트럼을 획득하는 프로브 타입의 근적외선 분광분석기보다는 적분구를 사용하거나 넓은 영역의 측정창을 사용하는 타입의 근적외선 분광분석기가 스펙트럼의 재현성 확보에 유리하다(Dahm and Dahm, 2003; Dahm and Dahm, 2004).

펄프 제지 산업계에서 근적외선을 이용한 공정제어 기술을 적극적으로 도입한 바 있으며(Wright *et al.*, 1990; Michell, 1995; Shimleck *et al.*, 1999; Raymond *et al.*, 2001), 목재의 주성분 함량 분석(Gierlinger *et al.*, 2002; Bergstrom, 2003; Poke *et al.*, 2004; Yeh *et al.*, 2005; Li *et al.*, 2007; Uner *et al.*, 2009; Uner *et al.*, 2011), 목재 내 수분 상태 분석(Tsuchikawa and Tsutsumi, 1998; Thygesen and Lundqvist, 2000; Eom *et al.*, 2010; Yang *et al.*, 2013; Yang *et al.*, 2015), 물리적 특성 분석(Shimleck *et al.*, 2003; Inakaki *et al.*, 2012; Alves *et al.*, 2012) 등 다양한 분야에서 연구가 진행되어왔다. 근적외선 영역에서 목재의 화학적 주성분들이 갖는 주요 흡광지점은 다음 Table 1-2와 같이 알려져 있다(Swanninger et al., 2011).

Table 1-2. Band assignments (from the literature and tentative assignments) in near-infrared region of wood and wood components. (Schwanninger et al., 2011)

Wavelength (nm)	Component	Bond vibration	Remarks
1143	Lignin	2^{nd} OT C_{ar}-H str. 2^{nd} OT C-H str. of CH_3 groups	CH_3 groups and aromatic moieties
1170	Lignin	2^{nd} OT asym. str. C-H, HC=CH	Lignin
1157 and 1171	Hemicellulose	2^{nd} OT C-H str.	CH_3 groups in acetyl ester groups in hemicelluloses
1188-1195	Lignin	2^{nd} OT C-H str.	CH_3 groups, lignin assigned form spectra
1212-1225	Cellulose	2^{nd} OT C-H str.	two to three bands t.a. CH and CH_2-groups, cellulose
1350	Hemicellulose	1^{st} OT C-H str. + C-H def.	Tentative assignment to CH_3 groups in acetyl ester groups in hemicelluloses and lignin and all wood components after acetylation
1360	All?	2^{nd} OT C-H str.	
1366	Cellulose	1^{st} OT C-H str. + C-H def.	Cellulose
1370	Hemicellulose	1^{st} OT C-H str. + C-H def.	Tentative assignment to CH_3 groups in acetyl ester groups in hemicelluloses and all wood components after acetylation
1386	-	1^{st} OT O-H str. 1^{st} OT C-H str. + C-H def.	isolated OH groups, C-H combination
1410	Lignin/ Extractives	1^{st} OT O-H str.	Phenolic hydroxyl groups or from lignin
1414	Water	1^{st} OT O-H str.	Formation of water-water H-bonds and development of water-water H-bonds depending on the water content in microcrystalline cellulose
1417	Lignin	1^{st} OT C-H str. + C-H bend.	aromatic associated C-H
1428	Cellulose/ H_2O	1^{st} OT O-H str. +H_20	Amorphous regions in cellulose + H_2O; free OH group or OH group with a weak H-bond, amorphous polysaccharides of wood, free and weakly H-bonded OH: O(6)-H(6) of cellulose, glucomannan; O(2)-H(2) of cellulose and xylan
1435-1438	-	1^{st} OT O-H str.	Water

Table 1-2. (Continued)

Wavelength (nm)	Component	Bond vibration	Remarks
1440	Lignin	1st OT C-H str. + C-H def.	CH or CH2- group
1447	Lignin/Extractives	1st OT O-H str.	Phenolic OH group
1448	Lignin	1st OT O-H str.	Phenolic groups of lignin with intramolecular H-bonding to an ether group in ortho position
1471	Hemicellulose	1st OT O-H str.	H-bonded O(6)-H(6) of glucomannan
1473	Cellulose	1st OT O-H str.	Semi-crystalline regions in cellulose
1476	Cellulose	1st OT O-H str.	Semi-crystalline regions in cellulose
1477-1484	Cellulose	1st OT O-H str.	Cellulose and absorbed water, weakly H-bonded OH of cellulose-O(6)-H(6)
1480	Cellulose	1st OT O-H str.	Semi-crystalline region in cellulose
1484-1493	Cellulose	1st OT O-H str.	Semi-crystalline region in cellulose
1489	Cellulose	1st OT O-H str.	Intramolecular H-bond in cellulose
1493	Hemicellulose	1st OT O-H str.	O(3)-H(3)···O(5) intramolecular H-bond of glucomannan
1510	Cellulose	1st OT O-H str.	O(2)-H(2) of cellulose (?)
1515	Cellulose?	1st OT O-H str.	O(6)-H(6)···O(3) interchain H-bond of cellulose
1534, 1550	Cellulose	1st OT O-H str.	H-bonded O-H groups of Cellulose
1540	Cellulose	1st OT O-H str.	Adsorption of Water
1545	Cellulose	1st OT O-H str.	Intramolecular H-bond in cellulose
1579	Cellulose	1st OT O-H str.	O(6)-H(6)···O(3) interchain H-bond of cellulose Strongly H-bonded O-H group in cellulose
1586-1596	Cellulose	1st OT O-H str.	Crystalline cellulose II
1588	Cellulose	1st OT O-H str.	Strong O(2)-H(2)···O(6) of cellulose
1591	Cellulose	1st OT O-H str.	H-bonds of the cellulose
1592	Cellulose	1st OT O-H str.	Crystalline regions in cellulose

Table 1-2. (Continued)

Wavelength (nm)	Component	Bond vibration	Remarks
1597	Cellulose	1st OT O-H str.	Strongly H-bonded O-H group in cellulose I, strongly H-bonded O-H groups in spruce wood
1616	-	1st OT C-H str.	$=CH_2$
1632	Cellulose	1st OT O-H str.	Band in microcrystalline cellulose and wood (found in bark tissue and compression wood, but no assignment)
1666	Hemicellulose	1st OT C-H str.	CH_3 groups
1668	Extractives	1st OT C_{ar}-H str.	Extractives
1672, 1674, 1677, 1673	Lignin	1st OT C_{ar}-H str.	Aromatic groups in lignin (aromatic skeletal due to lignin), aromatic C-H due to lignin
1681	Hemicellulose	1st OT C-H str.	CH_3 groups
1685	Lignin	1st OT C_{ar}-H str.	Aromatic ring associated
1695	-	1st OT C-H str.	CH_3 groups
1698	Lignin	1st OT C-H str.	C-H of lignin
1703	Cellulose	1st OT C-H str.	Cellulose, most probably from CH_2 groups
1705	Hemicellulose	1st OT C-H str.	CH_2 groups
1710	Hemicellulose	1st OT C-H str.	Furanose or pyranose due to hemicelluloses
1720	Lignin/ Hemicellulose/ (Cellulose)	1st OT C-H str.	All componenets, CH_3 groups
1724	Hemicellulose	1st OT C-H str.	Furanose/pyranose due to hemicellulose
1726	Lignin	1st OT C-H str.	C-H stretching of CH_2-groups
1731	Cellulose	1st OT C-H str.	
1765		1st OT C-H str.	CH_2-groups
1780	Cellulose	1st OT C-H_2 str.	Cellulose
1788 1790	Cellulose	1st OT C-H str.	Semi-crystalline or crystalline regions in cellulose
1791	Lignin	1st OT C-H str.	Appears in wood and milled wood lignin
1793	Cellulose	1st OT C-H str.	Cellulose

Table 1-2. (Continued)

Wavelength (nm)	Component	Bond vibration	Remarks
1811	Lignin	Unknown	Appears in wood and milled wood lignin
1820	Cellulose	O-H str. + 2^{nd} OT C-O str.	Cellulose?
1830	Cellulose	O-H str. + 2^{nd} OT C-O str.	Semi-crystalline or crystalline regions in cellulose
1907, 1910	Hemicellulose	2^{nd} OT C=O str.	Hemicellulose
1916-1942	Water	O-H str. + O-H def. of H_2O	Water
1980	Water	O-H str. + O-H def. of H_2	Water
2080	Cellulose	O-H str. + C-H def.	Semi-crystalline or crystalline regions in cellulose
2086	Cellulose/Hemi cellulose	O-H str. + O-H and C-H def.	Cellulose, xylan
2092	Cellulose	O-H and C-H def. + O-H str.	Cellulose and xylan
2110	Cellulose	O-H def. + O-H str.	Cellulose
2134	Lignin/ Extractives	C_{ar}-H str. + C=C str.	Lignin and extractives
2134	Hemicellulose	C-H str. + C=O str.	Acetyl groups in hemicellulose
2170, 2178	Cellulose/ Hemicellulose	Not assigned	Cellulose, xylan
2200	Lignin	C-H str. + C=O str.	Could be confirmed in wood and milled wood lignin
2255	All	O-H str. + C-O str. ???	Acetyl groups in acetylated wood
2267	Lignin	O-H str. + C-O str. ???	Lignin
2270	Cellulose	O-H str. + C-O str.	Cellulose
2271	Cellulose/ Hemicellulose	C-H_2 str. + C-H_2 def.	Cellulose and hemicellulose
2272	Hemicellulose	C-H str. + C-H def.	CH_3 groups
2277	Cellulose	O-H str. + C-C str. and/or C-H str. + C-H def.	Cellulose
2291	Cellulose	C-O str. + O-H str. or C-H_2 bend. + C-H_2 str.	Cellulose
2328-2332	Hemicellulose	C-H str. + C-H def.	Hemicelluloses, xylan

Table 1-2. (Continued)

Wavelength (nm)	Component	Bond vibration	Remarks
2335	Cellulose/ Hemicellulose	C-H str. + C-H def.	Cellulose, hemicelluloses, xylan
2336	Lignin	C-H str. + C-H_2 def. ??	Milled wood lignin hardwood and softwood, (tentative assignment)
2338	Cellulose	C-H str. + C-H_2 def.	Semi-crystalline or crystalline regions in cellulose
2343	Cellulose	C-H str. + C-H def. and/or 2^{nd} OT C-H def.	Cellulose
2352	Cellulose	C-H str. + C-H def. 2^{nd} OT C-H_2 bend.	Cellulose
2361	Cellulose	O-H def. or C-H def. + C-H str. or C-H_2 str.	Cellulose
2380	Holocellulose?	2^{nd} OT O-H def.	???
2384	Lignin	Not assigned	Lignin, could be confirmed with milled lignin
2461	Carbohydrate	C-H str. + C-C str.	Starch
2491	Cellulose	C-H str. + C-C str.	Cellulose ???
2488	Lignin	C-H str. + C-C str.??	Milled wood lignin hardwood and softwood

Notes) OT : overtone, sym : symmetric vibration, asym : asymmetric vibration,
str. : stretching vibration, bend : bending vibration, t.a. : tentative assignment

3.2. 다변량 데이터 분석

근적외선 스펙트럼과 같은 다차원의 변량을 가지는 데이터를 해석하고 활용하기 위해서는 다변량 통계 분석을 적용하는 것이 효과적이다. 다변량 통계 분석은 관측치들에 대해 조사된 변수간의 상관관계를 해석하는 방법으로 구조적 단순화, 군집화 및 분류, 의존성 분석, 시각화 등의 목적으로 시도되고 있다. 이 중, 근적외선 스펙트럼 데이터의 분석에는 데이터의 변화 양상을 파악하거나 군집 분석을 수행하기 위한 주성분 분석과 회귀 모델을 개발하기 위한 부분 최소 자승법이 주로 활용된다. 회귀 분석은 관측치들이 가지는 다차원의 독립변수와 종속변수 사이의 상관관계를 분석하여 모델화함으로써 관측치의 독립변수만으로 종속변수를 예측하는 식을 개발하는 것을 주목적으로 한다. 근적외선 스펙트럼을 이용한 다변량 회귀 분석은 근적외선 스펙트럼을 독립변수로 하여 물성간의 통계적인 상관관계를 모델화하여 새로이 측정되는 스펙트럼 데이터로부터 해당 물성을 예측하고자 하는 것이다. 근적외선과 같은 다차원 데이터를 이용한 분류 모델로는 주성분 분석에 기반한 Soft independent modelling of class analogy (SIMCA), 부분 최소 자승법에 기반한 부분 최소 자승 판별 분석(Partial least squares discriminant analysis) 등이 활용된다. 다변량 데이터는 최근 딥러닝 등으로 각광받고있는 신경망(Neural network)을 이용한 분류가 실시될 수 있다. 신경망은 순전파와 역전파를 반복하여 적합한 가중치를 학습하는 모델로서 데이터로부터 비선형적 상관관계를 모델화한다. 신경망을 이용한 분류 모델로는 인공신경망(Artificial neural network), 합성곱 신경망(Convolution neural network) 등이 활용된다. 본 연구는 근적외선 스펙트럼의 다변량 분석에 의한 수종 분류를 목적으로 한다. 이에 본 절에서는 본 연구에서 활용된 다변량 데이터를 이용한 분류법의 기초를 정리하였다.

3.2.1. 주성분 분석

Pearson (1901)에 의해 개발된 주성분 분석(Principal componenet analysis, PCA)은 고차원의 데이터를 구성하는 독립변수간의 직교대각화된 상관성을 추출함으로써 새로운 설명 변수를 생성해내고 이를 통해 데이터를 재구성하는 다변량 통계 분석법이다.

주성분 분석 과정을 기하학적으로 서술하면 다음과 같다. 데이터의 상관관계 분석이 용이하도록 다차원 데이터 공간에서 데이터의 분산이 최대가 되는 새로운 좌표축을 찾고, 이 축으로 정사영 되는 크기를 새로운 좌표 값이 되도록 하는 좌표 변환을 실시한다. 이때 정의된 새로운 축을 주성분(principal component, PC, $\boldsymbol{p}_i$)이라고 하며 이 축에 투영시켜 얻은 새 좌표를 그 축에 대한 score (t_i)라 한다. 이후, 변환된 데이터 공간에서 새로운 축과는 직교하는 축 중에 데이터의 분산을 최대로 하는 축 방향을 찾고 데이터를 해당 축으로 정사영하는 좌표 변환을 반복한다.

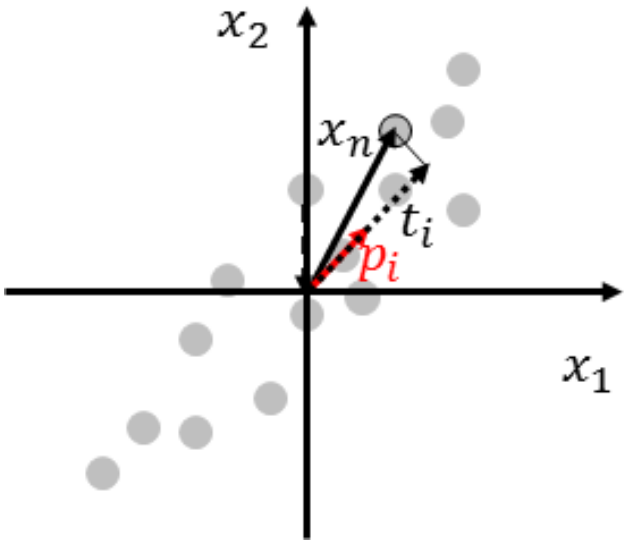


Figure 1-3. Geometric interpretation of principal component analysis in two dimensions.

주성분 분석의 실시 예를 2차원으로 설명하면 다음과 같다. Fig. 1-3과 같이 두 개의 차원을 가지는 데이터 벡터 $\boldsymbol{x}_n = (x_1, x_2)$이 N 개 존재할 때, 데이터의 분포를 1차원적으로 가장 잘 표현하는 방법은 데이터의 분산이 가장 큰 방향벡터, 즉 주성분 벡터 $\boldsymbol{p}_i$ 를 새 축으로 삼아 $\boldsymbol{p}_i$ 축으로의 정사영이 새로운 좌표(t_i)가 되도록 데이터를 변환하는 것이다. 이는 2차원 데이터 벡터를 1차원으로 감소시키면서도 원래 데이터의 정보를 크게 잃지 않는 차원 변환 방법이 된다.

이를 일반화하여 m 차원 데이터 벡터 $\boldsymbol{x}_n$의 분산을 최대로 하는 주성분 벡터 $\boldsymbol{p}_i$ 는 다음과 같이 얻을 수 있다. m 차원 데이터 벡터 $\boldsymbol{x}_n$이 N 개 (열방향 평균 = 0) 존재할 때, $\boldsymbol{x}_n$을 임의의 단위 벡터 $\boldsymbol{p}_i$ 로 정사영 한 크기 t_i 는 $\boldsymbol{x}_n$ 와 $\boldsymbol{p}_i$ 의 내적과 같다(Abdi and Williams, 2010).

$$t_i = \boldsymbol{x}_n \cdot \boldsymbol{p}_i \qquad \text{(Eq. I-1)}$$

이때, 전체 데이터 N 개의 정사영의 크기 t_i 의 분산 $\sigma_{p_i}^2$는 아래와 같이 표현된다.

$$\sigma_{p_i}^2 = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n \cdot \boldsymbol{p}_i)^2 - \left\{\frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n \cdot \boldsymbol{p}_i)\right\}^2 \qquad \text{(Eq. I-2)}$$

x_n의 평균이 0이므로, t_i의 평균의 제곱인 2번째 항 또한 0이 되므로 분산은 다음과 같이 간략화되며

$$\sigma_{p_i}^2 = \frac{1}{N}\sum_{n=1}^{N}(\boldsymbol{x}_n \cdot \boldsymbol{p}_i)^2 \qquad \text{(Eq. I-3)}$$

이를 행렬식으로 표현하면 아래와 같이 정리된다.

$$\begin{aligned}\sigma_{p_i}^2 &= \frac{1}{N}(\boldsymbol{XP})^T(\boldsymbol{XP}) \\ &= \frac{1}{N}\boldsymbol{P}^T\boldsymbol{X}^T\boldsymbol{XP} \\ &= \boldsymbol{P}^T\frac{\boldsymbol{X}^T\boldsymbol{X}}{N}\boldsymbol{P} \\ &= \boldsymbol{P}^T\boldsymbol{CP}\end{aligned} \qquad \text{(Eq. Ⅰ-4)}$$

여기에서 C 는 X 의 공분산 행렬(covariance matrix, $\boldsymbol{C} = cov(\boldsymbol{X})$을 의미한다. 이때, 분산의 최대값을 구하기 위한 조건은 $\boldsymbol{P}$ 가 단위벡터일 때 P^TCP 의 최대값을 구하는 문제이므로, 제약 등식(equality constraint)을 조건으로 하는 함수의 극대를 구하는 문제가 된다. 따라서 라그랑주 승수법을 적용하여 분산의 최대값의 조건을 구할 수 있으며(Klein, 2004), 다음 Eq. Ⅰ-5와 같이 최적화 방정식을 세울 수 있다.

$$u = \boldsymbol{P}^T\boldsymbol{CP} - \lambda(\boldsymbol{P}^T\boldsymbol{P} - 1) \qquad \text{(Eq. Ⅰ-5)}$$

이 때, C의 최대값은 $\frac{\partial u}{\partial P} = 0$ 일 때 존재하므로, Eq. Ⅰ-5에 $\boldsymbol{P}$ 에 대한 편미분을 실시하면 아래 식과 같이 정리된다.

$$\begin{aligned}&\frac{\partial \boldsymbol{u}}{\partial \boldsymbol{P}} = 2\boldsymbol{CP} - 2\lambda\boldsymbol{P} = 0 \\ &\therefore \boldsymbol{CP} = \lambda\boldsymbol{P}\end{aligned} \qquad \text{(Eq. Ⅰ-6)}$$

즉, 데이터 행렬 $\boldsymbol{X}$ 의 공분산 행렬 $\boldsymbol{C}$ 의 고유벡터(eigen vector)가 $\boldsymbol{X}$의 분산을 최대로 하는 방향벡터인 주성분 $\boldsymbol{p}_i$ 가 된다. 이로서 각 주성분 벡터는 데이터 $\boldsymbol{X}$ 의 분산의 방향성을 나타낸다. 데이터 $\boldsymbol{X}$의

주성분을 추출하는 방법은 결국 m 차원의 데이터 N 개가 존재하는 데이터 행렬 $\boldsymbol{X}$ $(m \leq N)$에 대하여, $\boldsymbol{X}$ 의 공분산 행렬 $\boldsymbol{C}$ 로부터 고유값(eigen value, λ_m)과 각각의 고유값에 대응하는 고유벡터인 주성분 $\boldsymbol{P}$ 를 계산하는 과정과 동일하다. 공분산행렬의 고유벡터와 고유값을 도출한다면, 정방대칭행렬인 공분산행렬은 항상 Eq. Ⅰ-8과 같이 대각화가 가능하다.

$$C \cdot \boldsymbol{p}_i = \lambda_i \cdot \boldsymbol{p}_i \ (i = 1, 2, \cdots, m) \qquad \text{(Eq. Ⅰ-7)}$$

$$C = \boldsymbol{P}\Lambda\boldsymbol{P}^T$$
$$= (\boldsymbol{p}_1 \cdots \boldsymbol{p}_m)\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_m \end{pmatrix}\begin{pmatrix} \boldsymbol{p}_1^T \\ \cdots \\ \boldsymbol{p}_m^T \end{pmatrix} \qquad \text{(Eq. Ⅰ-8)}$$

이렇게 얻어진 고유값의 크기에 따라 정렬($\lambda_1 > \lambda_2 > \cdots > \lambda_m$)한 고유벡터 $\boldsymbol{p}_i^T$는 i번째 주성분이 되며, 주성분 분석에서는 이를 loading 벡터라 칭한다. 데이터 행렬 $\boldsymbol{X}$ 와 고유벡터 $\boldsymbol{p}_i$ 간의 내적은 데이터 행렬 $\boldsymbol{X}$ 의 주성분 축으로의 정사영인 score를 의미하므로, score 벡터 $\boldsymbol{t}_i$는 다음과 같이 얻어진다.

$$\boldsymbol{t}_i = \boldsymbol{X} \cdot \boldsymbol{p}_i \ (i = 1,2,\cdots,m) \qquad \text{(Eq. Ⅰ-9)}$$

이를 행렬로 표현하면 다음 Eq. Ⅰ-10과 같으며,

$$\boldsymbol{T} = \boldsymbol{XP} \qquad \text{(Eq. Ⅰ-10)}$$

loading 벡터는 직교하므로 다음 Eq. Ⅰ-11이 성립하기 때문에,

$$\boldsymbol{P}^{-1} = \boldsymbol{P}^{T} \qquad \text{(Eq. I-11)}$$

Eq. I-10과 Eq. I-11을 조합하여 $\boldsymbol{X}$ 에 따라 정리하면 Eq. I-12와 같다.

$$\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}^{T} \qquad \text{(Eq. I-12)}$$

Eq. I-12를 이용하여 적절한 개수의 주성분을 정보량에 따라 선택하면 데이터의 차원을 감소시키면서도 원 데이터 $\boldsymbol{X}$ 의 정보를 크게 잃지 않은 상태로 아래 Eq I-13과 같이 압축($\boldsymbol{X}'$)할수 있다. ($\boldsymbol{E}$ 는 잔차)

$$\boldsymbol{X} = \boldsymbol{X}' + \boldsymbol{E} = (\boldsymbol{t}_1, \cdots, \boldsymbol{t}_i)\begin{pmatrix} \boldsymbol{p}_1^T \\ \cdots \\ \boldsymbol{p}_i^T \end{pmatrix} + \boldsymbol{E} \quad (i \leq m) \qquad \text{(Eq. I-13)}$$

이를 이용하면 데이터의 특성을 추출하고, 특성에 기반한 정보값인 score를 이용해 군집분석을 시행할 수 있다.

3.2.2. SIMCA (Soft independent modelling of class analogy)

SIMCA(Wold 1976)는 주성분 분석에 기반하여 실시하는 교사학습(Supervised learning)방법으로, 학습 데이터 내 각 집단(Class)에 독립적으로 주성분 분석을 실시함으로써 연산된 집단별 관측치의 잔차를 분포화하여 신뢰수준에 따른 분류를 실시한다. SIMCA에 의한 분류는 미지 시험편의 데이터를 각 집단의 주성분 분석 결과에 의해 재구성한 뒤 발생하는 잔차 값이 유의 수준 이내인지를 판단함으로써 분류를 실시한다.

전체 학습 데이터 $\boldsymbol{X}$ 중, 집단 K 의 데이터를 $\boldsymbol{X}_K$ (평균 ≠ 0)라 할 때, K 집단의 데이터만을 이용한 주성분 분석 결과는 Eq. Ⅰ-13을 변형하여 다음 Eq. Ⅰ-14와 같이 나타낼 수 있다.

$$\boldsymbol{X_K} = \boldsymbol{X_K}' + \boldsymbol{E}_K = \overline{\boldsymbol{X_K}} + \boldsymbol{T}_K \boldsymbol{P}_K^T + \boldsymbol{E}_K \qquad \text{(Eq. Ⅰ-14)}$$

($\overline{\boldsymbol{X}_K} = \boldsymbol{X}_K$의 열 방향 평균 행렬)

이 때, 고유값의 크기대로 선택된 r 개의 주성분에 의해 잔차($\boldsymbol{E}_K$)가 결정된다. 주성분 분석은 잔차($\boldsymbol{E}_k$)를 계속하여 데이터의 차원(m)까지 분해할 수 있으므로, 잔차는 $(m-r)$개의 주성분으로 구성된다. 집단 K 내 관측치 k의 i번째 독립변수에서의 잔차를 e_{ki}^K라 할 때, 집단 K를 구성하는 모든 관측치가 갖는 잔차의 표준편차 s_o는 아래 Eq. Ⅰ-15와 같다. 이것으로 학습집단 내 데이터들의 r개 주성분을 포함하는 공간에서의 분포를 결정할 수 있다(Maesschalck et al., 1999).

$$s_o = \sqrt{\sum_{k=1}^{N_K}\sum_{i=1}^{m}[(e_{ki}^K)^2/(m \bullet N_K)]} \qquad \text{(Eq. Ⅰ-15)}$$

미지 시험편의 데이터 $\boldsymbol{x}_j^{new}$ 가 집단 K 에 속하는지 여부를 판단하는 과정은 다음과 같다. 미지 시험편의 데이터 $\boldsymbol{x}_j^{new}$ 를 집단 K의 주성분 분석 결과로 선택된 r개의 주성분으로 정사영하여 각 주성분마다의 score로 구성된 벡터인 $\boldsymbol{t}_j^{new}$를 다음 Eq. Ⅰ-16에 따라 연산한 후, 이를 Eq. Ⅰ-16과 같이 r 개의 score($\boldsymbol{t}_j^{new}$)와 loading으로 재구성($\boldsymbol{x'}_j^{new}$)한다.

$$\begin{aligned} \boldsymbol{t}_j^{new} &= [\boldsymbol{x}_j^{new} - \overline{\boldsymbol{x}_K}]\boldsymbol{P}_K \\ \boldsymbol{x'}_j^{new} &= \overline{\boldsymbol{x}_K} + \boldsymbol{t}_j^{new}\boldsymbol{P}_K^T \end{aligned} \qquad \text{(Eq. Ⅰ-16)}$$

이렇게 재구성된 관측치 $\boldsymbol{x'}_j^{new}$와 실제 관측치 $\boldsymbol{x}_j^{new}$간의 차를 이용하여 집단 K 주성분 공간(주성분 개수 = r)에 의해 미지 시험편이 갖는 잔차 $\boldsymbol{e}_j^{new}$는 다음과 같다.

$$\boldsymbol{e}_j^{new} = \boldsymbol{x}_j^{new} - \boldsymbol{x'}_j^{new} \qquad \text{(Eq. Ⅰ-17)}$$

이에 따라 미지 시험편이 집단 K 주성분 공간에서 갖는 잔차의 표준편차 s_j는 아래와 같다.

$$s_j = \sqrt{\sum_{i=1}^{m}(e_{ji}^{new})^2/m} \qquad \text{(Eq. Ⅰ-18)}$$

집단 K 주성분 공간에 의한 잔차 들이 정규분포를 따른다고 가정할 때, 미지 시험편의 학습 집단 소속 여부를 결정하는 방법은 자유도가 m인 미지 시험편과 자유도가 $m \bullet N_K$ 인 학습 집단의 잔차의 분산

이 유의수준 α에 의한 F-검정으로 차이를 보이는지를 검증함으로써 가능하다. 해당 자유도와 유의수준 α에서의 F값을 F_{crit}이라고 할 때, 유의적으로 동일한 분산의 경계치 s^2_{crit}는 아래와 같다.

$$s^2_{crit} = F_{crit} s^2_0 \qquad \text{(Eq. I-19)}$$

미지 시험편 잔차의 분산 s^2_j 와 집단 K 가 갖는 잔차의 분산 s^2_0이 학습 집단 K 의 분포 내에 유의적으로 포함되어 있는지를 판단하는 방법, 즉 집단 K 로 분류할 수 있는지 여부를 판단하는 방법은, F-검정에 의해 다음과 같이 계산되는 F^{new}_j 를 이용하여 판단할 수 있다.

$$F^{new}_j = \frac{s^2_j}{s^2_o} \qquad \text{(Eq. I-20)}$$

$$\begin{aligned} &\text{if} F^{new}_j \le F_{crit}, \quad x^{new}_i \in class\, K \\ &\text{if} F^{new}_j > F_{crit}, \quad x^{new}_i \notin class\, K \end{aligned} \qquad \text{(Eq. I-21)}$$

SIMCA는 학습 집단별 주성분 공간에서의 잔차 분포를 이용하여 미지 시험편의 잔차가 집단별 잔차의 분포에 포함되는지 여부를 유의수준에 따라 판단함으로써 분류를 실시하는 알고리즘으로 볼 수 있다.

3.2.3. 부분 최소 자승법

회귀분석은 종속변수 데이터 Y 를 독립변수 데이터 X 로 설명하는 방법이다. m차원의 독립변수 x_i $(i = 1, 2, \cdots, m)$와 1차원의 종속변수 y 사이의 선형적인 상관관계는 다음의 Eq. Ⅰ-22로 표현할 수 있다.

$$y = b_0 + x_1 b_1 + x_2 b_2 + \cdots + x_m b_m + f \qquad \text{(Eq. Ⅰ-22)}$$

상기 Eq. Ⅰ-22에서 b_i를 회귀계수라 하며, f는 오차 또는 잔차라 한다. 상기 식은 아래와 같이 정리하여 Eq. Ⅰ-24와 같이 행렬화할 수 있다.

$$y = \sum_{i=1}^{m} x_i b_i \qquad \text{(Eq. Ⅰ-23)}$$

$$Y = XB + F \qquad \text{(Eq. Ⅰ-24)}$$

독립변수 데이터 X $(N \times M)$와 종속변수 데이터 Y $(N \times 1)$간의 상호관계를 선형적으로 설명하는 회귀계수 B 는 최소자승법(least squares)을 이용하여 구할 수 있으며, 이는 Eq. Ⅰ-25와 같다. 이를 다변량 선형 회귀분석 (Multivariable linear regression, MLR)이라 한다. 이를 이용하면 독립변수와 종속변수간의 상관관계가 회귀계수 B 에 의해 모델화된다.

$$B = (X^T X)^{-1} X^T Y \qquad \text{(Eq. Ⅰ-25)}$$

MLR에 의해 회귀계수를 구하기 위해서는 $X^T X$의 역행렬이 반드시 존재해야한다. 그러나 실제적으로는 대개 두 가지 요인에 의해 $X^T X$의 역행렬이 존재하지 않는다. 첫째로, 독립변수의 차원(m)이 관측치의 개수(n) 보다 큰 경우($n < m$), 회귀계수 B 가 무한개로 존재한다. 이를 극

복하여 최적의 유일한 B 를 구하기 위해서는 독립변수를 배제하는 방법이 있다. 그러나 최적의 회귀계수를 도출하기 위한 독립변수 배제 기준의 설정이 어렵다. 둘째로, $n \geq m$ 이나, 독립변수들이 종속적인 관계를 갖고 있어 X^TX의 행렬식(determinant)이 0이 되는 경우이다. 이 경우 작은 오차 요인에 의해 여러 개의 독립변수가 동시에 영향 받게 되므로 노이즈에 의한 종속변수 추정의 정확도가 크게 낮아진다. 다변량 선형 회귀분석이 가진 두 가지 문제점을 해결하는 방법은, 독립변수 X가 Full rank를 갖는 행렬이 되도록 재구성하는 것이다(Geladi and Kowalski, 1986).

부분 최소 자승법(partial least squares, PLS)은 상기와 같은 문제가 해결되는 회귀분석법으로, 종속도가 높은 데이터에 대해 높은 선형 회귀 성능을 나타낸다. 주성분 분석을 확장하여 X 데이터 뿐 아니라 Y 데이터에도 주성분 분석을 실시하여 다변량 회귀분석을 실시한다. 부분 최소 자승법은 근사된 데이터 공간을 이용하여 회귀분석을 수행하는 것으로, 외적 관계(X 및 Y block 각각)와 내적 관계(두 block간의 연결, score 벡터간의 회귀분석)로 구성되어있다고 볼 수 있다. 데이터의 관측 수가 n 개이고, 독립변수가 m 개이며, 이에 따른 종속변수가 M 개일 때, X는 $(n \times m)$ 행렬이 되며, Y 는 $(n \times M)$ 행렬이 된다.

X와 Y간의 상관관계를 최대로 하는 요인 $\boldsymbol{w}$(주성분 분석에서의 주성분)을 추출하기 위해 Y^TX의 공분산으로부터 고유벡터를 찾으면 아래와 같다(Lindgren et al., 1993).

$$C_{Y^TX} = cov(Y^TX)$$
$$C_{Y^TX}\boldsymbol{w} = \lambda \boldsymbol{w}$$

$$C_{Y^TX} = \boldsymbol{W}\boldsymbol{L}\boldsymbol{W}^T = (\boldsymbol{w}_1 \cdots \boldsymbol{w}_m)\begin{pmatrix} l_1 & & \\ & \ddots & \\ & & l_m \end{pmatrix}\begin{pmatrix} \boldsymbol{w}_1^T \\ \cdots \\ \boldsymbol{w}_m^T \end{pmatrix} \quad \text{(Eq. I-26)}$$

요인 W 에 의한 X의 정사영인 T^*는 다음과 같이 정의되며, 동일한

방법으로 X^TY의 공분산으로 추출한 요인 C에 의한 Y의 정사영인 U^*와 Y는 다음과 같다.

$$T^* = XW$$
$$U^* = YC \qquad \text{(Eq. I-27)}$$

이를 이용하여 T^*와 U^*간의 최소자승법에 의한 β를 다음과 같이 구할 수 있다.

$$U^* = T^*\beta + G$$
$$\beta = (T^{*T}T^*)^{-1}T^{*T}U^* \qquad \text{(Eq. I-28)}$$

Eq. I-27과 Eq. I-28를 이용하여 다음과 같은 선형 회귀 방정식을 구할 수 있다.

$$U^* = T^*\beta + G$$
$$YC = XW\beta + G \qquad \text{(Eq. I-29)}$$
$$Y = XW\beta C^T + GC^T$$
$$Y = XB + F \qquad \text{(Eq. I-30)}$$

따라서 부분 최소 자승법은 근사화된 X 데이터와 Y 데이터간의 상관관계를 선형적으로 도출한 것이 되면서, 오류 F 를 최소화한 형태가 된다. 부분 최소 자승법에 의한 회귀 모델이 포함하는 주성분은 주성분 분석에서의 주성분과 구분하기 위해 일반적으로 factor 또는 latent variable으로 명명한다. 부분최소자승법은 독립변수와 종속변수간의 선형적인 상관관계가 존재할 때 독립변수간의 상관도가 높은 데이터에서도 예측 신뢰도가 높은 모델을 개발할 수 있다는 장점이 있다.

3.2.4 부분 최소 자승 판별 분석

부분 최소 자승법을 이용한 분류 방법인 부분 최소 자승 판별 분석(Partial least squares discriminant analysis)은 종속변수인 Y 데이터에 1 또는 0으로 정하는 class value를 대입하는 분류 방법이다. 부분 최소 자승 판별 분석은 전체 데이터 X 중, 집단 K 의 스펙트럼 데이터를 X_K , 종속변수를 Y_K 라 하고, 이 외 집단의 스펙트럼 데이터를 X_{notK} , 종속변수를 Y_{notK} 라 할 때, $Y_K = 1$, $Y_{notK} = 0$ 으로 하는 부분 최소 자승 모델을 개발하는 것이다. 이 때 개발된 모델의 Y 값이 기준치 이상이면 집단 K 에 속하는 것으로, 이외의 경우는 K 집단이 아닌 것으로 판정한다.

미지의 스펙트럼이 소속될 집단의 판별을 위한 기준치를 결정하기 위해서는 예측된 Y 값을 이용한다. 가장 단순한 수준의 판별은 i번째 데이터 X_i에 대하여, 판별 모델에 의해 예측된 $y_{i,predicted}$ 값을 그대로 활용하는 방법으로, 집단 K 의 종속변수인 1과 이외 집단의 종속변수인 0의 중간 값인 0.5를 기준으로 한다. 즉, $y_{i,predicted}$가 0.5 이상이면 집단 K 에 속하는 것으로, 반대로 0.5 미만이면 집단 K 에 속하지 않는 것으로 분류하는 방법으로, 아래의 식과 같이 표현된다.

$$\begin{aligned} X_i &\subset K \quad \text{if } y_{i,predicted} \geq 0.5 \\ X_i &\subset notK \text{ if } y_{i,predicted} < 0.5 \end{aligned} \qquad \text{(Eq. I-31)}$$

K 개의 다중 집단에 대한 분류를 시행하는 문제에서는, 하나의 관측치 데이터가 K 개의 집단을 판별하기 위한 모델에 각각 대입되므로 K 개의 $y_{i,predicted}$ 가 도출되며, 이를 이용하여 집단 판별을 실시한다. 따라서 단순히 0.5를 기준치로 판별을 실시할 수 없다. K 개의 $y_{i,predicted}$를 이용하여 집단 판별 분석을 실시하는 방법 중, 가장 엄격한 수준의 판별 방법은 0.5 이상인 $y_{i,predicted}$가 단 1개 집단에서만 존재할 때, 관측치를 해

당 집단으로 판정하고 이외의 경우는 판별을 기각하는 것이다.

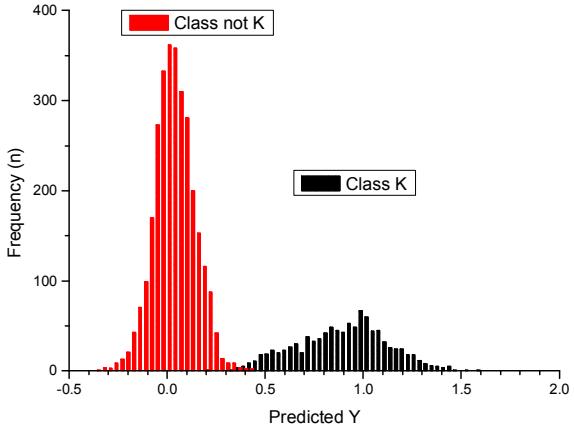


Figure 1-4. Histogram of predicted Y of cross validation by partial least squares discriminant analysis.

Fig. 1-4와 같이 집단 K 와 집단 $not\,K$ 를 구분하는 판별 모델의 예측 결과 $y_{i,predicted}$을 이용하여 도수분포를 분석하면 다음과 같은 특징을 확인할 수 있다. (1) 집단 K 판별 모델에 의해 집단 K 및 집단 $not\,K$ 관측치들이 갖는 $y_{i,predicted}$ 값의 중심 값이 각각 정확히 1과 0이 아니다. (2) 0.5를 기준으로 경계를 구분할 때, 집단 K 내 관측치 중 상당수의 $y_{i,predicted}$ 값이 0.5 미만이다. 이들은 집단 K 임에도 불구하고 집단 K 의 판별 모델에 따라 집단 $not\,K$ 로 판정되었다 (false negative). (3) 집단 $not\,K$ 내 관측치들은 대부분 0.4 미만의 $y_{i,predicted}$ 를 가지고 있었다.

따라서 $y_{i,predicted}$ = 0.5 가 항상 가장 유용한 기준치가 되지는 않으며, 목적에 따라 기준치를 적절히 재조정하면 모델의 회귀계수와 예측값들을 그대로 유지하면서도 판별 성능만을 개선시킬 수 있다. 이를 달성하기 위해 분포를 이용한 확률적 접근을 적용할 수 있다. Fig. 1-4의 집단별 도수분포는 $y_{i,predicted}$를 이용하여 정규 분포를 형성하는 확률밀도함수(probability density fuction)로서 간주될 수 있

다. 집단 K 가 갖는 예측 결과의 평균을 $y_{predicted,mean,K}$, 표준편차를 σ_K라 할 때 집단 K 와 집단 $notK$ 의 예측 결과가 가지는 도수분포는 다음에 식에 따라 각각 정규 분포를 따르는 확률 밀도 함수 f 로 표현할 수 있다(Perez et al., 2009).

$$f(y_{predicted,K}) = \frac{1}{\sigma_K \sqrt{2\pi}} e^{-(y_{predicted,K} - y_{predicted,mean,K})^2 / 2\sigma_K} \quad \text{(Eq. I-32)}$$

$$f(y_{predicted,notK}) = \frac{1}{\sigma_{notK} \sqrt{2\pi}} e^{-(y_{predicted,notK} - y_{predicted,mean,K})^2 / 2\sigma_{notK}} \quad \text{(Eq. I-33)}$$

집단 K 판별 모델에 의해 단일 관측치 x_i가 갖는 종속변수 $y_{i,predicted}$ 가 집단 K 에 속할 확률 $F(y_{i,predicted} \subset K)$는 확률 밀도 함수 $f(y_{predicted,K})$의 구간 $[-\infty, y_{i,predicted}]$에서의 면적이 되며, 반대로 집단 $notK$ 에 속할 확률 $F(y_{i,predicted} \subset notK)$은 확률 밀도 함수 $f(y_{predicted,notK})$의 구간 $[y_{i,predicted}, \infty]$에서의 면적이 된다.

$$F(y_{i,predicted} \subset K) = \int_{-\infty}^{y_{i,predicted,K}} f(y_{predicted,K}) \quad \text{(Eq. I-34)}$$

$$F(y_{i,predicted} \subset notK) = \int_{y_{i,predicted,K}}^{\infty} f(y_{predicted,notK}) \quad \text{(Eq. I-35)}$$

모든 $y_{i,predicted}$ 는 집단 K 이거나 집단 $notK$에 속하는 2가지 경우만 존재하기 때문에, 집단 $notK$ 일 확률의 합이 1이 되도록 정규화가 필요하며, 정규화된 확률 F_{norm} 는 다음과 같다(Pérez *et al.*, 2009).

$$F_{norm}(y_{i,predicted} \subset K) = \frac{F(y_{i,predicted} \subset K)}{F(y_{i,predicted} \subset K) + F(y_{i,predicted} \subset not\,K)} \quad \text{(Eq. I-34)}$$

$$F_{norm}(y_{i,predicted} \subset not\,K) = \frac{F(y_{i,predicted} \subset not\,K)}{F(y_{i,predicted} \subset K) + F(y_{i,predicted} \subset not\,K)} \quad \text{(Eq. I-35)}$$

따라서 F_{norm}을 이용하면 $y_{i,predicted}$ 는 집단 K 에 속할 확률(0 ~ 1)로 변환됨으로써 다음과 Eq. I-36을 이용하여 집단 K 에 속할 확률이 0.5 이상일 때를 기준으로 판별을 실시할 수 있도록 한다.

$$\begin{array}{ll} X_i \subset K & \text{if } F_{norm}(y_{i,predicted}) \geq 0.5 \\ X_i \subset not\,K & \text{if } F_{norm}(y_{i,predicted}) < 0.5 \end{array} \quad \text{(Eq. I-36)}$$

또한, 이는 집단 K 에 속할 확률과 집단 $not\,K$에 속할 확률이 Eq. I-37과 같이 동일하도록 하는 y_i를 찾는 과정으로 볼 수 있다.

$$F_{norm}(y_i \subset K) = F_{norm}(y_i \subset not\,K) = 0.5 \quad \text{(Eq. I-37)}$$

이를 만족하는 y_i를 새로운 경계치인 $y_{criteria}$ 라 할 때 Eq. I-36은 Eq. I-31을 변형하여 다음 Eq. I-38과 같이 표현할 수 있다.

$$\begin{array}{ll} X_i \subset K & \text{if } y_i \geq y_{criteria} \\ X_i \subset not\,K & \text{if } y_i < y_{criteria} \end{array} \quad \text{(Eq. I-38)}$$

다중 분류 문제에서의 단일 관측치에 의하여 발생하는 K 개의 F_{norm} 값을 이용하여 판별을 수행하는 여러 방법이 있으나, 가장 엄격한 수준의 판별 방법은 0.5 이상인 F_{norm}가 단 1개 집단에서만 존재할 때,

관측치를 해당 집단으로 판정하고 이외의 모든 경우는 판별을 기각하는 것이다. 이때 발생하는 판별 기각에는 2가지 종류로 나뉜다. (1) 0.5 이상의 확률을 나타낸 집단이 존재하지 않아 발생한 판별 기각. (2) 0.5 이상의 확률을 나타낸 집단이 2개 이상 존재하여 발생한 판별 기각. 전자의 경우를 미분류라 하며, 후자의 경우를 중복 분류라 할 수 있다.

3.2.5. 신경망 이론

신경망(Neural network) 이론은 단일 신경세포의 신호전달을 해석하기 위한 McCulloch와 Pitts(1943)의 인공 신경 모델링에서부터 시작하였다. 이를 기반으로 인공신경망을 구축하면 인간의 지능을 모사한 인공지능을 구현하여 패턴 인식 등이 가능함이 알려지면서 전 세계적으로 활발한 연구가 진행되었다. 이후 60년대에 이르러 당시의 컴퓨터로는 대규모의 인공신경망 계산 효율성이 낮다는 점과 Minsky와 Papert(1969)에 의해 인공신경망이 XOR 분류를 학습할 수 없다는 점이 밝혀지면서 인공신경망을 이용한 연구가 한계에 부딪혔다. 이후 Webos(1974)가 제안한 오차 역전파에 의한 학습 알고리즘이 1986년 Rumelhart 등에 의해 대중화되면서, 인공신경망의 정확도와 학습 범주가 확장되었다. 현대에 이르러 인공신경망은 컴퓨터의 연산 능력이 급속도로 발전함에 따라 빅데이터를 활용한 딥러닝 기법으로 발전하였다.

McCulloch와 Pitts(1943)는 뉴런의 신호 전달을 해석하기 위한 계산 모델인 인공신경 모델을 최초로 제안하였다. 이들이 제안한 인공신경 모델은 임계값 (u) 이상의 신호만을 하부 뉴런으로 전달하는 특징을 수학적으로 모델화한 것으로, 인공신경으로 들어오는 n개의 입력 신호(x_i, $i = 1, 2, \cdots, n$)의 가중합(weighted(w_i) sum)이 임계치(u)를 초과하였을 때 하부로 전달되는 출력 신호(y)는 1이 되고, 그렇지 않은 경우는 0이 되는 단위계단 함수(θ)이다.

$$y = \theta\left(\sum_{i=1}^{n} w_i x_i - u\right) \qquad \text{(Eq. I-39)}$$

하나의 인공 신경은 Fig. 1-5와 같이 입력(Input), 활성화 함수(Activation function) 및 출력(Output) 세 가지 요소로 구성되어 있다. 입력은 인공 신경으로 유입되는 신호의 크기를 의미한다. 활성화 함수는 인공 신경으로 입력된 신호의 가중합을 하단으로 전달할 신호의 크기를

결정하는 함수로서 연구 개발 초기에는 임계치 이상의 신호를 전달하는 단위 계단 함수 (unit step function)가 주로 사용되었으나, 이후 활용에 따라 신호함수(signum), 선형 함수(linear fuction), 시그모이드 함수(sigmoid function) 및 쌍곡 함수(hyperbolic function) 등이 활성화 함수로 활용되고 있다(Fig. 1-6). 출력은 다음 인공신경으로 전달하는 신호 크기를 의미한다.

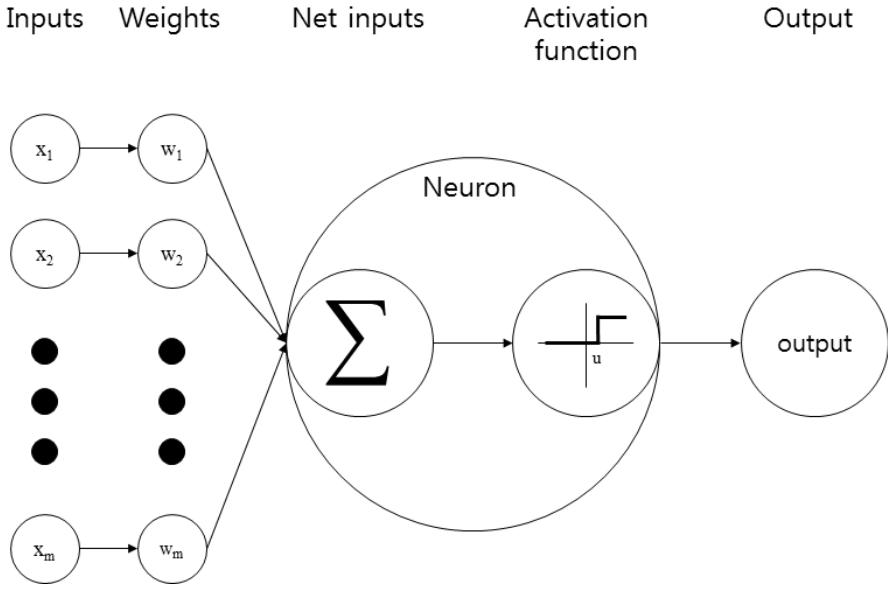


Figure 1-5. Signal transduction structure of a single artificial neuron.

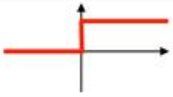
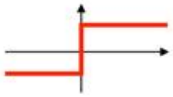
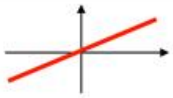
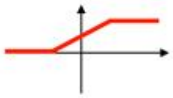
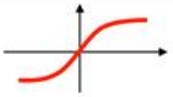
Activation function	Equation	Example	1D Graph
Unit step (Heaviside)	$\phi(z) = \begin{cases} 0, & z < 0, \\ 0.5, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Sign (Signum)	$\phi(z) = \begin{cases} -1, & z < 0, \\ 0, & z = 0, \\ 1, & z > 0, \end{cases}$	Perceptron variant	
Linear	$\phi(z) = z$	Adaline, linear regression	
Piece-wise linear	$\phi(z) = \begin{cases} 1, & z \geq \frac{1}{2}, \\ z + \frac{1}{2}, & -\frac{1}{2} < z < \frac{1}{2}, \\ 0, & z \leq -\frac{1}{2}, \end{cases}$	Support vector machine	
Logistic (sigmoid)	$\phi(z) = \frac{1}{1 + e^{-z}}$	Logistic regression, Multi-layer NN	
Hyperbolic tangent	$\phi(z) = \frac{e^{z} - e^{-z}}{e^{z} + e^{-z}}$	Multi-layer NN	

Figure 1-6. Most common activation functions.

인공신경망은 상기와 같은 인공신경을 층에 따라 독립적으로 배열하고 각 층을 병렬적으로 연결시켜 데이터에 대한 학습을 수행한다. 다른 층으로의 신호 전달은 신호 전달의 강도를 나타내는 가중치에 따라 결정된다. 인공신경망의 가장 기초적인 모델은 세 개의 층으로 구성되며 각각 입력층(Input layer), 은닉층(Hidden layer) 및 출력층(Output layer)으로 구성된다. 이를 도식화하면 Fig. 1-7과 같다.

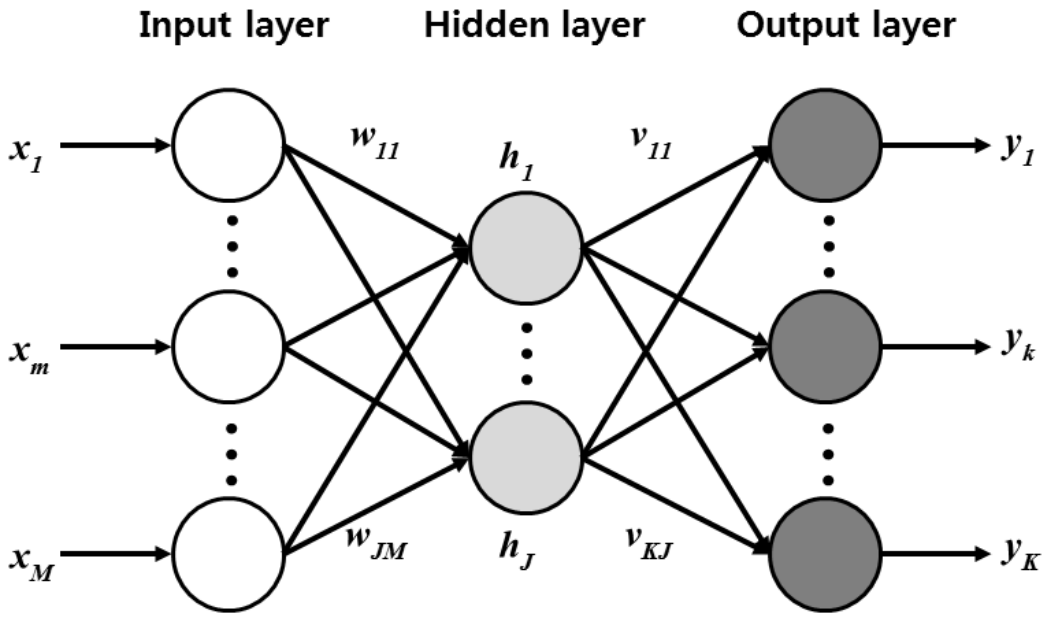


Figure 1-7. Artificial neural network. (3 layer)

입력층의 노드 수(= 입력 데이터의 차원 수)를 M, 은닉층의 노드 수를 J, 출력층의 노드 수(= 예측 값의 종류 또는 집단 수)가 K 이고, 입력층 내 m번째 노드와 은닉층 내 j번째 노드 사이의 가중치를 w_{jm}, 은닉층 내 j번째 노드의 bias를 b_j라고 하면, 은닉층 내 j번째 노드의 입력 p_j는 다음과 같다.

$$p_j = \sum_{m=1}^{M} (w_{jm} x_m + b_j) \qquad \text{(Eq. I-40)}$$

은닉층의 활성화 함수 f에 의해, 은닉층 내 j번째 노드의 출력 h_j 는 다음과 같다.

$$h_j = f(p_j) \qquad \text{(Eq. I-41)}$$

이를 벡터화하면 다음과 같이 표현된다.

$$\boldsymbol{p} = \boldsymbol{W}\boldsymbol{x}^T + \boldsymbol{b}_j \quad \text{(Eq. I-42)}$$
$$\boldsymbol{p} : (J \times 1),\ W : (J \times M),\ \boldsymbol{x}^T : (M \times 1),\ \boldsymbol{b} : (J \times 1)$$

$$\boldsymbol{h} = f(\boldsymbol{p}) \quad \text{(Eq. I-43)}$$
$$\boldsymbol{h} : (J \times 1)$$

이와 마찬가지로, 은닉층 내 j번째 노드와 출력층의 k번째 노드 사이의 가중치를 v_{kj}, 출력층 내 k번째 노드의 bias를 c_k라고 하면, 출력층 내 k번째 노드의 입력 q_k는 다음과 같다.

$$q_k = \sum_{j=1}^{J}(v_{kj}h_j + c_k) \quad \text{(Eq. I-44)}$$

출력층의 활성화 함수 g에 의해, 출력층의 k번째 노드에서의 출력 y_k 는 다음과 같다.

$$y_k = g(q_k) \quad \text{(Eq. I-45)}$$

이를 벡터화하면 다음과 같이 표현된다.

$$\boldsymbol{q} = \boldsymbol{V}\boldsymbol{h} + \boldsymbol{c} \quad \text{(Eq. I-46)}$$
$$\boldsymbol{q} : (K \times 1),\ V : (K \times J),\ \boldsymbol{h} : (J \times 1),\ \boldsymbol{c} : (K \times 1)$$

$$\boldsymbol{y}^T = g(\boldsymbol{q}) \quad \text{(Eq. I-47)}$$
$$\boldsymbol{y}^T : (K \times 1),\ \boldsymbol{q} : (K \times 1)$$

상기와 같은 구조를 갖는 인공신경망을 이용한 학습은 N 개의 M 차원의 데이터 X^T $(M \times N)$를 투입하여 필요로 하는 출력값 Y^T $(K \times N)$를 찾도록 하는 가중치(W, V)와 bias(b, c)를 구하는 과정을

의미한다.

인공신경망이 가중치를 학습하는 알고리즘은 입력층으로부터 출력층으로 신호가 전달되는 순전파(feed forward)와 출력층으로부터 입력층까지 오차를 전달하는 역전파(back propagation)의 두 가지 방향으로 실시된다. 인공신경망 연구 초기에는 가중치에 대한 학습을 위해 순전파를 실시하였다. 그러나 인공신경망 연구 초기의 단방향적 학습 알고리즘은 각 층을 구성하는 노드간의 가중치 정보가 복잡해질수록 학습이 어려웠다(Jain et al., 1996).

이러한 단점은 Rumelhart (1986)에 의해 역전파 알고리즘이 신경망의 가중치를 효과적으로 학습시킬 수 있다는 점이 조망되며 극복되었다. 역전파는 미분 가능한 활성화 함수를 갖는 인공신경망의 경우에서 적용 가능한데, 인공신경망의 출력과 종속변수 간의 차이인 오차를 출력층으로부터 입력층으로 분배하는 방식으로 가중치 및 bias를 수정하는 알고리즘이다(Wythoff, 1993; Whitley, 1995)

데이터의 독립변수 X^T $(M \times N)$를 인공신경망의 입력층에 투입하여 초기 가중치 $\boldsymbol{w}_j, \boldsymbol{v}_k$ 및 bias $\boldsymbol{b}_j, \boldsymbol{c}_k$에 의해 순전파되어 계산된 출력층의 출력값을 $Y^T_{predicted}$ $(K \times N)$라 할 때, n 번째 관측치에 대한 오차 E_n 은 순전파 과정 중에 사용된 가중치 $\boldsymbol{w}_j, \boldsymbol{v}_k$ 및 bias b, c에 의해 결정되므로, 각각에 인자에 대한 E_n으로의 편미분을 연쇄법칙을 이용하여 전개하면 다음과 같다(Ibraheem 2013).

$$\frac{\partial E_n}{\partial \boldsymbol{w_j}} = \frac{\partial E_n}{\partial p_j}\frac{\partial p_j}{\partial \boldsymbol{w}_j} = \frac{\partial E_n}{\partial p_j}\boldsymbol{x} \qquad \text{(Eq. I -48)}$$

$$\frac{\partial E_n}{\partial b_j} = \frac{\partial E_n}{\partial p_j}\frac{\partial p_j}{\partial b_j} = \frac{\partial E_n}{\partial p_j} \qquad \text{(Eq. I -49)}$$

$$\frac{\partial E_n}{\partial \boldsymbol{v}_k} = \frac{\partial E_n}{\partial q_k}\frac{\partial q_k}{\partial \boldsymbol{v}_k} = \frac{\partial E_n}{\partial q_k}\boldsymbol{h} \qquad \text{(Eq. I -50)}$$

$$\frac{\partial E_n}{\partial c_k} = \frac{\partial E_n}{\partial q_k}\frac{\partial q_k}{\partial c_k} = \frac{\partial E_n}{\partial q_k} \qquad \text{(Eq. I -51)}$$

따라서, $\frac{\partial E_n}{\partial p_j} = \delta_j$와 $\frac{\partial E_n}{\partial q_k} = \delta_k$를 연산하면 모든 인자들이 오차에 미치는 영향을 계산할 수 있다.

$$\delta_k = \frac{\partial E_n}{\partial q_k} \qquad \text{(Eq. I -52)}$$

$$\begin{aligned}
\delta_j = \frac{\partial E_n}{\partial p_j} &= \sum_{k=1}^{K}\frac{\partial E_n}{\partial q_k}\frac{\partial q_k}{\partial p_j} \\
&= \sum_{k=1}^{K}\delta_k\frac{\partial(v_{kj}h_j + c_j)}{\partial p_j} \\
&= \sum_{k=1}^{K}\delta_k\frac{\partial(v_{kj}f(p_j) + c_j)}{\partial p_j} \\
&= \sum_{k=1}^{K}\delta_k(f'(p_j)v_{kj}) \\
&= f'(p_j)\sum_{k=1}^{K}\delta_k v_{kj}
\end{aligned} \qquad \text{(Eq. I -53)}$$

δ_k와 δ_j는 각각의 노드가 입력받은 오차라고 볼 수 있으므로, 은닉층의 노드들이 가지는 오차 δ_j는 출력층의 노드들이 갖는 오차 δ_k와 출력층과 입력층 사이의 가중치 v_{kj}의 가중합에 의해 순전파의 반대방향으로 전달

된다.

최종적으로 Eq. Ⅰ-48과 Eq. Ⅰ-50에 의하여 경사하강법(Gradient descent)에 의한 인공신경망의 학습을 실시하는 경우, 초기 가중치 $\boldsymbol{v}_k$와 $\boldsymbol{w}_j$를 학습률(learning rate) η 만큼 갱신한 가중치 $\boldsymbol{v}_{k,new}$, $\boldsymbol{w}_{j,new}$는 다음과 같다.

$$\boldsymbol{v}_{k,new} = \boldsymbol{v}_k - \eta\delta_k\boldsymbol{h} \qquad \text{(Eq. Ⅰ-54)}$$

$$\boldsymbol{w}_{j,new} = \boldsymbol{w}_j - \eta\delta_j\boldsymbol{x} \qquad \text{(Eq. Ⅰ-55)}$$

인공신경망의 출력 노드에 각 관측치가 갖는 class value (0 또는 1)를 넣는 경우, 인공신경망은 분류를 위한 학습을 수행하게 된다.

분류 알고리즘에서는 출력층의 출력 y_k을 확률로 변환하기 위하여 베이즈 정리(Bayes theorem)에 기반한 분류를 수행한다(Richard and Lippmann 1991; Zhang 2000). 베이즈 정리는 두 확률 변수의 사전 확률과 사후 확률 사이의 관계를 나타내는 것으로, 베이즈 정리(Eq. Ⅰ-56)에 의해 사전 확률로부터 사후 확률을 구할 수 있다.

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \qquad \text{(Eq. Ⅰ-56)}$$

데이터 분류의 관점에서, $P(Y|X)$는 데이터 X가 주어졌을 때 집단 Y에 속할 사후 확률(Posterior), $P(X|Y)$는 각 집단에 속해있는 데이터 X의 확률 분포인 우도(Likelihood), $P(Y)$는 집단들의 분포에 대한 사전 확률(Prior), $P(X)$는 데이터의 분포를 의미한다. 전체 데이터 X가 K개의 집단으로 구성되어 있다면, $P(X)$는 다음 Eq. Ⅰ-57과 같이 쓸 수 있다.

$$P(X) = \sum_{k=1}^{K} P(X|Y_k)P(Y_k) \qquad \text{(Eq. Ⅰ-57)}$$

연산의 용이성을 확보하기 위하여 다음 Eq. Ⅰ-58와 같이 자연로그를 이용하여 a_k를 정의하면 Eq Ⅰ-58과 같으며, 이를 이용하여 Eq. Ⅰ-56을 전개하면 데이터 X가 집단 k에 속할 확률은 Eq. Ⅰ-59와 같이 변형된다.

$$a_k = \ln\left(P(X|Y_k)P(Y_k)\right) \qquad \text{(Eq. Ⅰ-58)}$$

$$P(Y_k|X) = \frac{P(X|Y_k)P(Y_k)}{\sum_{k=1}^{K} P(X|Y_k)P(Y_k)} = \frac{e^{a_k}}{\sum_{k=1}^{K} e^{a_k}} \qquad \text{(Eq. Ⅰ-59)}$$

따라서 데이터 X가 다중 집단 중 집단 k에 속할 확률을 우도와 사전 분포에 의해 결정할 수 있다. 일반적인 다중 집단의 분류에서 우도는 대개 정규 분포(Gaussian distribution)를 따르는 것으로 알려져 있다 (Akaike 1998; Penny and Roberts, 1999; Banerjee et al., 2008). 그러나 정규분포함수는 미분이 어려우며, 연산이 복잡하다. 이를 단순화하기 위해 사전 확률과 우도를 구분하지 않고, Eq. Ⅰ-44의 q_k를 입력하여 회귀하는 문제로 단순화하면, 정규분포의 누적확률함수와 매우 흡사한 형태를 띄고 있으며, 미분이 용이한 특징을 가지는 softmax 함수(Eq. Ⅰ-60)로 근사된다. softmax 함수의 미분은 자기 자신으로 구성된 값을 나타내는 특징이 있어 오차의 역전파에서 연산을 단순화해주는 장점이 있다 (Williams and Barber 1998).

$$P(Y_k|X) = y_k = softmax(q_k) = \frac{e^{q_k}}{\sum_{k=1}^{K} e^{q_k}} \quad \text{(Eq. I-60)}$$

다중 분류 학습을 위해 신경망 이론에서 사용하는 오차 E_n는 최우추정(Maximum likelihood estimation)에 기반하여 정의된다(Bishop 2006). 학습 세트에 N개의 데이터(x_n)들이 K개의 집단에 속해있을 때, 실제 속한 집단의 성분만 1, 이외의 성분은 0으로 하는 정답 데이터를 벡터화 ($\boldsymbol{t}_n$)하여 $\boldsymbol{t}_n$의 o번째 성분인 t_{no}는 다음과 같이 정의할 수 있으며, 이를 one-hot encoding이라고 한다.

$$y_{measured} := t_{no} = \begin{Bmatrix} 1 \ \text{if} o = k \\ 0 \ \text{if} o \neq k \end{Bmatrix} \quad \text{(Eq. I-61)}$$

N 개의 데이터가 K 개의 집단에 대해 확률적으로 얼마나 잘 분류되었는지를 평가하는 방법은 각 데이터의 사후확률을 곱하는 것으로 다음과 같이 정의되며, 이를 우도함수(Likelihood function)라고 한다.

$$L(X) = \prod_{n=1}^{N}\prod_{k=1}^{K} P(Y_k|x_n)^{t_{nk}} = \prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{nk}} \quad \text{(Eq. I-62)}$$

우도함수 $L(X)$는 학습 데이터가 인공신경망의 인자인 W, V, b, c에 의해 출력된 y_{nk}가 가장 적합한 확률, 즉 t_{nk}에 가까울 때 최대값을 가질 것이므로, $L(X)$의 최대값을 갖도록 하는 W, V, b, c를 찾아야하며 이 과정을 최우추정이라 한다. $L(X)$의 최대값은 $L(X)$의 미분값이 0일 때 존

재할 것이나, 곱연산의 형태인 $L(X)$는 미분이 복잡하므로 미분이 용이한 형태로 변환하기 위해 양변에 로그를 취하여 합연산으로 변환하고, 부호를 반전한 $E(X)$를 정의하면 다음과 같다.

$$E(X) = -\ln(L(X)) = -\ln(\prod_{n=1}^{N}\prod_{k=1}^{K} y_{nk}^{t_{nk}}) = -\sum_{n=1}^{N}\sum_{k=1}^{K}(t_{nk}\ln y_{nk}) \quad \text{(Eq. I-63)}$$

이 때 정의된 $E(X)$를 cross entropy 함수라 정의하며 Loss라 칭한다. $L(X)$의 값을 최대로 하는 W, V, b, c는 마찬가지로 Loss ($E(X)$)를 최소로 하는 인자가 될 것이므로, 신경망 이론에서는 Loss를 오차로 정의하고 오차가 최소가 되도록 하는(예측확률이 최대인) 인자들을 찾는 최적화를 오차의 역전파에 의해 실시한다.

신경망에 학습을 위한 최적화 방법 중 앞서 언급한 경사하강법(Gradient descent)이 가진 느린 학습속도, 학습률이 높은 경우 발산하는 문제, 전역 최소값이 아닌 지역 최소값으로 수렴하는 문제 등을 해결하기 위해 다양한 최적화 방법들이 개발되었다(Stochastic gradient descent (Robbins and Monro, 1951), Momentum gradient descent (Rumelhart et al., 1986), Nesterov Momentum (Sutskever et al., 2013), Adagrad (Duchi et al., 2011), Adadelta (Zeiler 2012), Adam (Kingma and Ba, 2014)).

이에 따라 초기 임의로 설정된 가중치로부터 순전파와 역전파를 반복하며 오차를 최소로 하는 가중치들을 찾아나갈 수 있다. 부분 최소 자승법이 선형적인 상관관계를 학습하는데 적합하다면, 비선형 활성화함수가 포함된 인공신경망은 비선형적 상관관계를 학습할 수 있는 능력을 갖고 있다.

합성곱 신경망(Convolution neural network, CNN)은 주로 화상 분류를 기반으로 한 분야에서 혁신적인 성능을 나타내고 있는 신경망 구조의 하나이다. 합성곱 신경망이 각광받게 된 것은 Lecun 등(1998)에 의해 발표된 문자 인식을 위한 합성곱 신경망이 오차 역전파와 경사하강법에 의한 최적화 학습에 의해 뛰어난 분류 성능을 나타냈다는 논문에서부터 시작되었다. ImageNet과 같은 분류 대회에서 합성곱 신경망 구조는 이미지를 이용한 분류 성능을 기존에 비해 크게 향상시켰으며(Krizhevsky et al., 2012; Zhou et al., 2014; Ren et al., 2014; Szegedy et al., 2015), 이외의 분류 문제에서 가장 널리 사용되는 인공신경망의 하나이다. Fig. 1-8에 Lecun 등(1998)이 발표한 2차원 합성곱 신경망의 구조를 나타내었다.

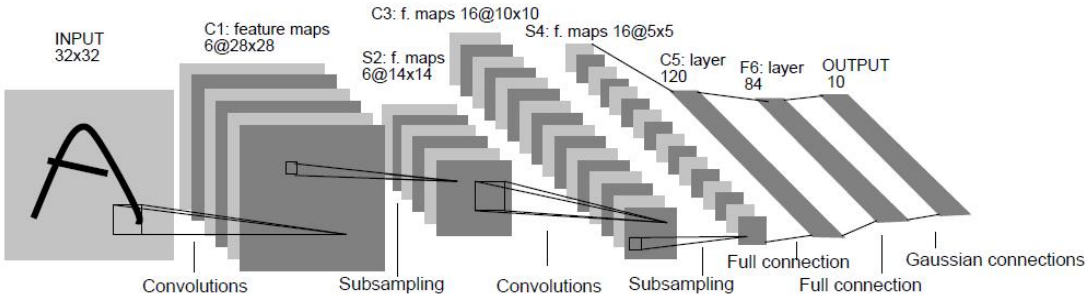


Figure 1-8. Architecture of LeNet-5, a Convolution Neural Network, here for digits recognition. Each plane is feature map. (Lecun et al., 1998)

이들이 제안한 합성곱 신경망은 합성곱 층(Convolution layer), 활성화 함수, 풀링층(Pooling layer, subsampling)이 반복되면서 후단에서 이미지 데이터를 일렬로 늘인 후 인공 신경망에서와 동일한 모든 노드가 연결되는 전연결층(Fully connected layer), 그리고 출력층으로 구성되어있다.

합성곱 층에서는 앞 층에서의 데이터를 필터와의 합성곱 연산에 의해 입력받는다. 합성곱 연산에 사용되는 $k_1 \times k_2$ 크기의 2차원 필터(또는 커

널) $W(k_1,k_2)$와 2차원 입력 데이터 X 사이에 정의되는 합성곱 연산에 의해 구성되는 합성곱층 좌표에 따른 값은 다음과 같다. 여기에서 b_{ij}는 bias를 의미한다.

$$O_{ij} = \sum_{m=0}^{k_1-1} \sum_{n=0}^{k_2-1} (X(i+m,j+n)\,W(m,n) + b_{ij}) \qquad \text{(Eq. I-64)}$$

이를 상세하게 설명하면, Fig. 1-9와 같이 3×3 크기의 입력 데이터 X와 2×2 크기의 필터 W에 의해 구성된 합성곱 데이터 O 는 다음과 같으며, 이는 입력 데이터 위에 필터를 얹고 동일한 위치에서의 데이터 값과 필터 값을 곱하여 합한 것과 같다.

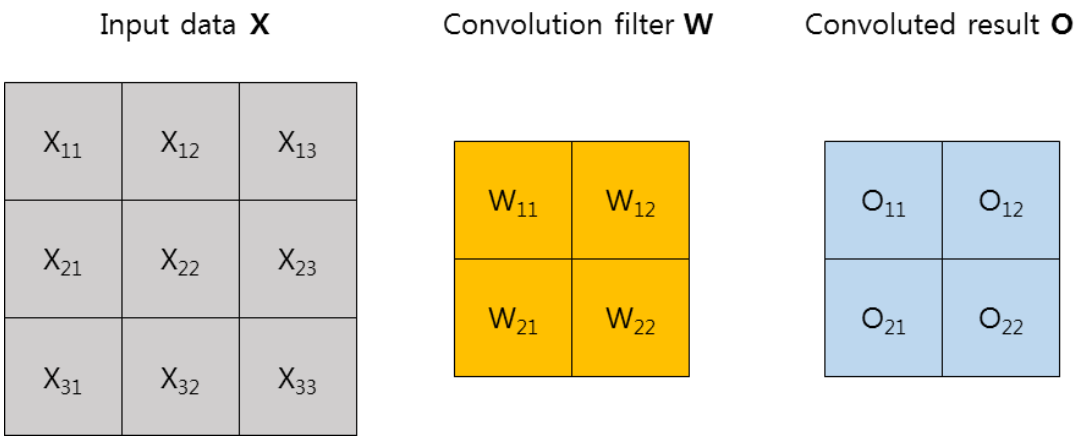


Figure 1-9. Example of Convolution calculation.

$$\begin{aligned}
O_{11} &= X_{11}W_{11} + X_{12}W_{12} + X_{21}W_{21} + X_{22}W_{22} + b_{11} \\
O_{12} &= X_{12}W_{11} + X_{13}W_{12} + X_{22}W_{21} + X_{23}W_{22} + b_{12} \\
O_{21} &= X_{21}W_{11} + X_{22}W_{12} + X_{31}W_{21} + X_{32}W_{22} + b_{21} \\
O_{22} &= X_{22}W_{11} + X_{23}W_{12} + X_{32}W_{21} + X_{33}W_{22} + b_{22}
\end{aligned} \qquad \text{(Eq. I-65)}$$

상기와 같은 연산을 이용하여 합성곱 필터는 2차원 데이터(주로 화상)

로부터 위치별로 특징을 추출하는 역할을 담당한다. 합성곱에는 이러한 필터의 개수에 따라 다양한 종류의 특징을 추출할 수 있다.

합성곱 신경망은 일반적으로 2차원 화상 데이터를 이용한 분류에서 유효한 성능을 나타내고 있다. 그러나, 본 연구에서는 1차원 데이터 벡터인 스펙트럼 데이터를 다루고 있으므로, 1차원 합성곱 신경망에 대해 중점적으로 다루고자 한다. 1차원 합성곱 신경망은 1차원 데이터 벡터에 합성곱을 실시하는 것을 의미하는 것으로, 1차원 합성곱 연산은 Eq. Ⅰ-64를 1차원으로 감소시켜 변형하는 것과 같다.

$$O_i = \sum_{m=0}^{k-1} (X(i+m)\, W(m) + b_i) \qquad \text{(Eq. Ⅰ-66)}$$

1차원 합성곱 신경망의 구조를 1개의 입력층, 1개의 합성곱층과 1개의 출력층으로 간략화한 예를 들면 Fig. 1-10과 같다(Acquarelli et al., 2017). 입력 데이터에 대해서 k크기를 가지는 필터에 의해 인접한 데이터간의 합성곱을 실시하여 데이터를 변형시킨다. 합성곱 연산 결과는 인공신경망에서 활용하였던 활성화 함수에 의해 활성화되고, 출력층과는 인공신경망과 동일하게 완전연결된 가중치들에 의해 신호를 보내어 분류학습을 실시한다. 1차원 합성곱 신경망은 이를 통해 가중치와 필터에 대한 학습을 수행한다. 1차원 합성곱 신경망에서는 필터가 학습된다는 점이 종래의 다변량 분석과는 큰 차이가 있다.

기존의 근적외선 스펙트럼을 이용한 다변량 분석에서는 스펙트럼 데이터의 적절한 수학적 전처리가 최종 모델의 성능에 크게 영향한다. 그러나, 1차원 합성곱 신경망에서는 입력 데이터와 출력층에 맞는 필터(가중치)를 목적에 맞게 스스로 학습한다. 즉 분류를 목적으로 하는 경우, 적절한 필터의 형태를 스스로 개발하여 분류에 최적화된 형태로 스펙트럼을 가공하는 신경망을 구성한다.

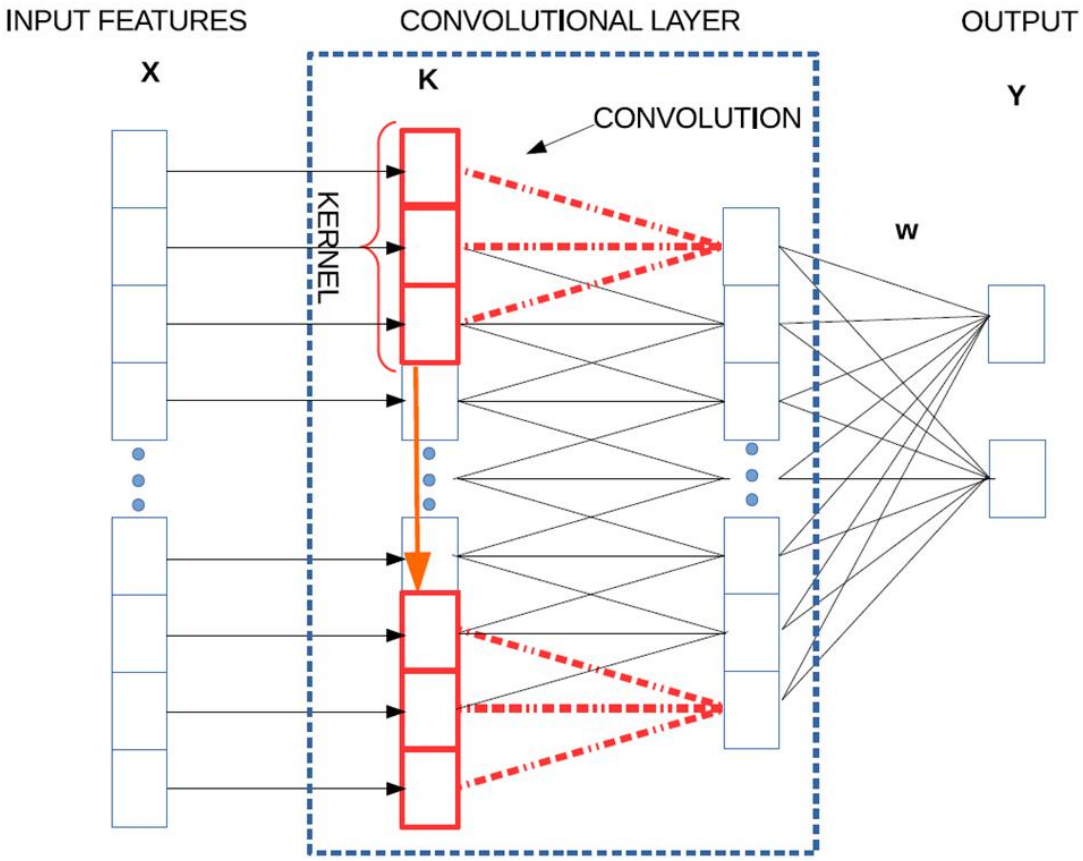


Figure 1-10. Example of 1 dimensional convolution neural network for a two class classification problem with one convolution layer and two output nodes. X means the input, K means the kernel(filter), w means weights and Y means the output (predicted class). (Acquarelli et al., 2017)

앞서 서술한 분류 알고리즘들(SIMCA, 부분 최소 자승 판별 분석, 인공 신경망, 1차원 합성곱 신경망)의 분류 인자, 판별 기준 및 장·단점을 정리하면 Table 1-3과 같다.

Table 1-3. Overview of classification methods developed in this study.

Classification Method	SIMCA	Partial least squares discriminant analysis	Artificial Neural Network	1 dimensional convolution neural network
Classification factor	• Residuals of each class principal component analysis	• $y_{predicted}$ values of each species regression model	• $y_{predicted}$ values of output layer	• $y_{predicted}$ values of output layer
Discrimination criteria	• Significance (α)	• $y_{criteria}$	• $y_{criteria}$	• $y_{criteria}$
Pros	• Provide spectral information	• Provide spectral information • Provide probability	• Self-adaptive modelling • Provide probability	• Self-adaptive modelling • Provide probability
Cons	• Depends on the number of principal components • Provide no probability	• Depends on a number of latent variables • Sensitive to outlier	• Provide no information of parameters • High computational cost • Needs large dataset	• Provide no information of parameters • High computational cost • Needs large dataset

4. 공시재료

본 연구에서는 국산 침엽수종을 구분하기 위해 활용된 수종은 총 다섯 수종으로 낙엽송(Larch, *Larix kaempferi*), 소나무(Red pine, *Pinus densiflora*), 잣나무(Korean pine, *Pinus koraiensis*), 삼나무(Cedar, *Cryptomeria japonica*), 편백(Cypress, *Chamaecyparis obtusa*)을 선정하였다. 상기 다섯 수종은 국내 제재목 생산업에 공급되는 국산 침엽수 원목 중 대다수를 차지한다(산림청, 2017).

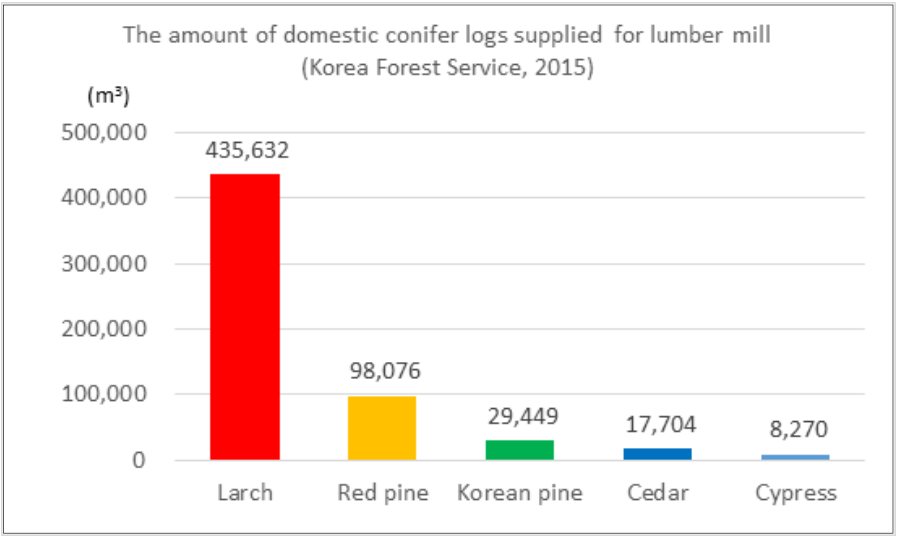


Figure 1-11. The amount of domestic conifer logs supplied for lumber mill (Korea Forest Service, 2015).

모든 시험편은 국내 여러 지역의 산림조합(산림조합 중부 목재 유통센터, 동부 목재 유통센터, 나주시 산림조합, 남원시 산림조합, 가평군 산림조합, 서귀포시 산림조합)에서 크기 50 × 100 × 600 mm (Thickness × width × length) 크기의 생재 상태의 제재목을 수종별로 50개씩 수집하였다. 수집된 생재 제재목은 온도 25℃, 상대습도 50 ~ 70%를 유지하는 공간에서 기건(함수율 10 ~ 15%)하였다. 기건이 종료된 후, 제재목의 넓은면을 대패가공 한 후 근적외선 스펙트럼을 측정하였다.

5. 근적외선 스펙트럼 측정

시험편의 기건이 완료된 후 공시재료로부터 근적외선 분광분석기(Spectra Star 2600XT-R, Unity Scientific, US)를 사용하여 목재 표면의 근적외선 스펙트럼을 측정하였다.

Figure 1-12. Near infrared spectrometer.

본 장비는 680 nm부터 2600 nm까지 파장대를 1 nm 해상도로 스펙트럼을 측정한다. 광원으로는 텅스텐 할로겐 램프를 사용하며, InGaAs 소자의 검출기가 설치되어있다. 광섬유를 이용한 프로브 타입과는 달리, 직사각형의 측정창(25 × 40 mm)에 시험편을 밀착하여 근적외선 스펙트럼을 측정한다. 이로 인해 프로브 타입에서 발생할 수 있는 이방성/비균질재료의 국소부위 신호에 따른 근적외선 측정 오차를 낮출 수 있다.

본 근적외선 분광분석기는 단색화장치가 포함된 pre-dispersive 타입으로, 장치의 하단에 설치된 광원으로부터 시험편에 단색광을 조사하고, 확산반사된 광을 45° 각도에서 측정한다. 반사율 측정은 장비의 기계적 잡음을 감소시키기 위해 한 시험편에서 12번 측정한 스펙트럼의 평균값을 사용하였다(scans to average = 12). 시험편의 반사율(R_m)은 표준 물질의 반사광($R_{ref,m}$) 대비 시험편의 반사광($R_{s,m}$)의 비율로써 Eq. Ⅰ-67과 같이 정의된다.

$$R_m = \frac{R_{s,m}}{R_{ref,m}} \times 100\,(\%) \qquad \text{(Eq. Ⅰ-67)}$$

근적외선 분광분석기를 이용하여 측정한 시험편의 반사율(R_m)은 Eq. Ⅰ-68을 이용하여 흡광도(A_m)로 변환하여 근적외선 영역의 흡광도 스펙트럼을 구성한다.

$$A_m = \log(1/R_m) \qquad \text{(Eq. Ⅰ-68)}$$

대패 가공된 각 제재목 시험편의 넓은 면에서 옹이 또는 결함이 존재하지 않는 심재부 5구역에서 근적외선 흡광도 스펙트럼을 측정한 후(Fig. 1-13), 약 2 mm를 대패 가공하였다. 이 과정을 4회 반복하여 각 시험편마다 20개의 근적외선 스펙트럼을 수집하였다. 따라서 각 수종별로 1,000개의 근적외선 스펙트럼 데이터가 수집되었으며, 총 다섯 수종에서 5,000개의 근적외선 스펙트럼이 수집되었다.

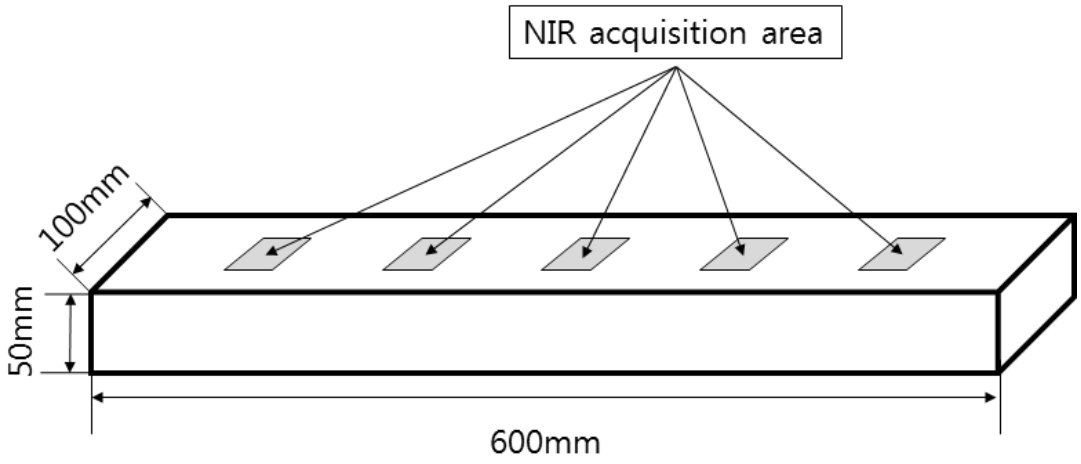


Figure 1-13. Schematic diagram of near-infrared spectrum acquisition from lumber specimen.

근적외선 대역에서는 작용기에 의한 산란과 중첩이 광범위하게 발생하므로, 스펙트럼을 수학적으로 변형하면 예측모델의 신뢰성을 높일 수 있다. 이러한 스펙트럼의 수학적 변형을 수학적 전처리(mathematical preprocessing)라 한다. 수학적 전처리 방법으로는 평활화(smoothing), 기준값 보정(baseline correction), 정규화(normalize), 미분(derivatives) 등이 널리 활용되며, 모델 개발의 신뢰성 개선을 위해 다양한 조합을 연속적으로 적용할 수 있다. 그러나, 과도한 수학적 전처리는 스펙트럼이 갖는 특성을 훼손할 수 있으므로 항상 모델의 신뢰성을 증대시킬 수 있는 것은 아니다. 따라서 근적외선 분광분석시, 스펙트럼의 적절한 수학적 전처리가 필요하다(Siesler et al., 2008).

본 연구에서는 수학적 전처리에 따른 분류 성능을 비교하여 최적조건을 선정하기 위하여 2가지의 상태의 수학적 전처리가 실시된 근적외선 스펙트럼을 이용한 분류 성능을 평가하였다 첫 번째는 대조군으로 원(raw) 스펙트럼들(Fig. 1-14)을 그대로 이용한다. 두 번째는 기준선 제거 (baseline correction) 및 산란 보정(scattering correction) 효과가 있는 standard normal variate (SNV)(Barnes et al., 1989)전처리를 실시(Fig. 1-15)하였다. SNV 전처리(Eq. Ⅰ-69)는 제재목과 같은 고형

시료의 표면 형상과 비균질성에 따라 발생할 수 있는 기준선의 차이 및 산란효과를 효과적으로 보정한다. 세 번째는 기준선 제거 및 흡광지점 분해 효과가 있는 Savitzky-Golay 2nd derivative(SG 2nd)을 polynomial order =2, window size = 21로 실시하였다(Fig. 1-16)(Savitzky and Golay, 1964). SG 2nd는 미분을 실시함으로써 기준선을 제거하는 효과를 갖고 있다. 또한 흡광영역이 중첩되어있는 근적외선 대역에서 효과적으로 흡광지점들을 분리하며, 강한 흡광지점들을 강조하는 효과를 가지기 때문에 근적외선 분광분석분야에서 널리 활용되고 있다.

$$x_m^{new} = \frac{x_m^{old} - \sum_{k=1}^{M}(x_k/M)}{\sigma(x_m^{old})} \qquad \text{(Eq. I-68)}$$

여기에서, x_m^{new} : m번째 파장에서의 새로운 스펙트럼 값

x_m^{old} : m번째 파장에서의 원래 스펙트럼 값

σ : 데이터의 표준편차

M : 데이터의 차원 수

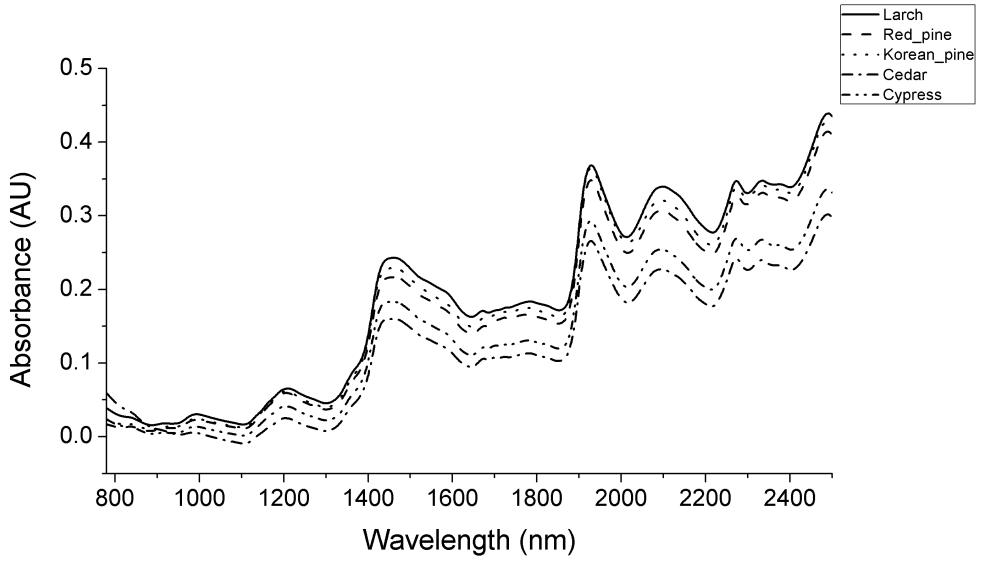


Figure 1-14. Raw near-infrared absorbance spectra of each species.

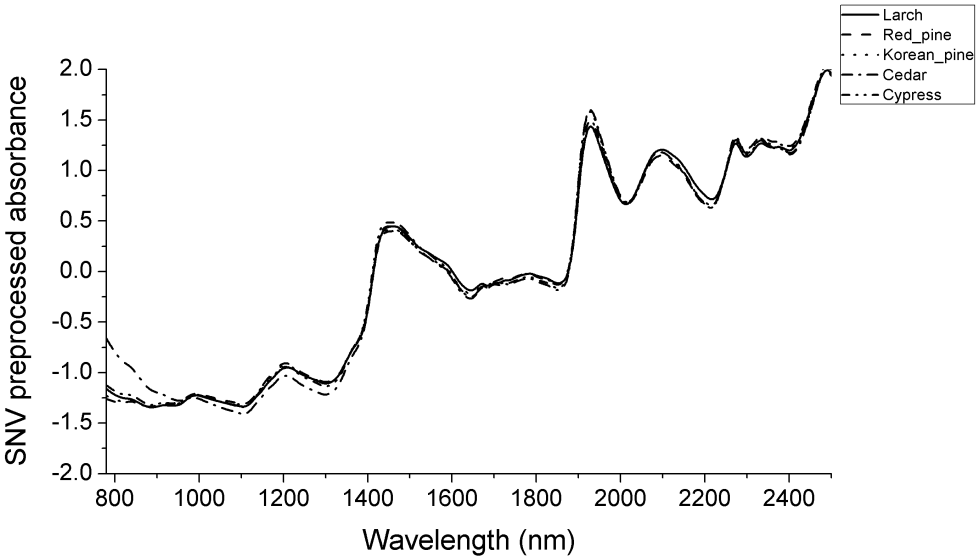


Figure 1-15. Standard normal variate preprocessed near-infrared absorbance spectra of each species.

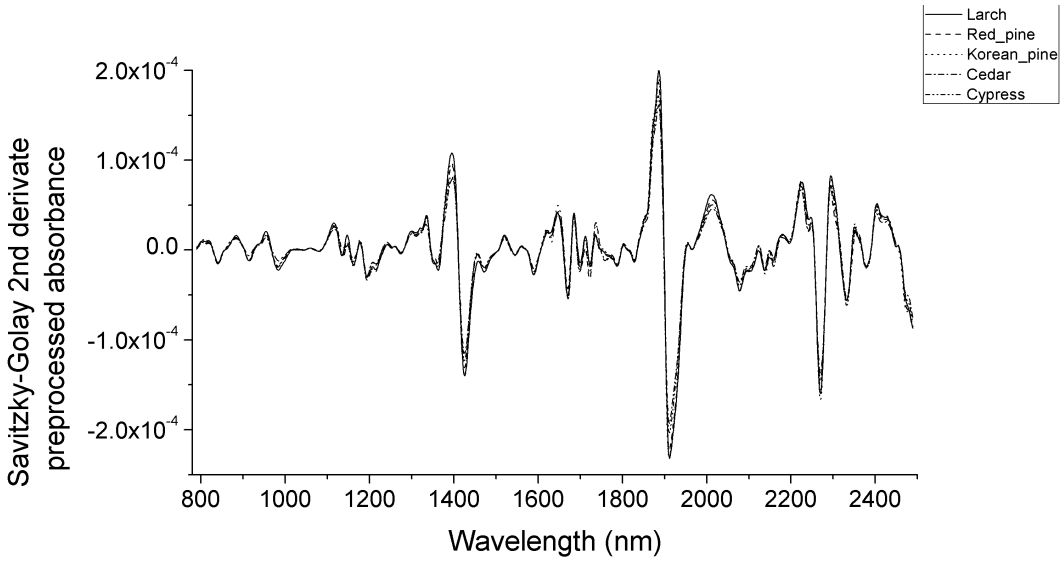


Figure 1-16. Savitzky-Golay 2nd derivative preprocessed spectra of each species.

제 2장

주성분 분석과 SIMCA를 이용한 수종 구분

1. 서론

근적외선 분광분석법은 화학적 성분에 대한 정보를 위주로 검출하므로, 근적외선 스펙트럼을 이용한 화학적 성분의 함량 또는 특성 분석이 가능하다. 목재는 크게 셀룰로오스, 헤미셀룰로오스, 리그닌 및 추출물과 같은 다양한 화학적 성분으로 구성된 세포가 집적된 이방성 재료다. 목재의 경우 수종에 따라 화학적 성분의 구성이 달라지므로, 이를 근거로 한 수종 구분에 관한 연구가 실시되었다. 근적외선 스펙트럼의 주성분 분석은 수집된 스펙트럼의 변화 양상을 분석하는 방법이므로, 충분한 수의 근적외선 스펙트럼을 확보하였을 때 수종 구분이 가능하다. 근적외선 스펙트럼을 이용한 목재 수종 구분에 관한 연구는 1990년대부터 시작되었으며, Brunner 등(1996)과 Shimleck 등(1996)에 의해 주성분 분석에 의한 군집분석이 시도되었다. 이들의 연구는 일련의 물리적(분말화)/화학적 처리에 의했을 때 수종 구분이 가능하다는 것으로, 초기 연구 단계에서는 실용적인 활용이 어려웠다. 이후 목재 칩(Michell and Shimleck 1998, Russ et al., 2009), 소시편(Flaete et al., 2006, Sandak et al., 2011, Nisgoski et al. 2017), 제재목(Adedipe et al., 2008) 등 다양한 형태의 목재에서 획득한 근적외선 스펙트럼으로 주성분 분석을 실시하여 군집화 분석 또는 soft independent modelling of class analogy (SIMCA)에 의한 수종 구분을 실시하였다. Adedipe 등(2008)은 SIMCA에 의해 red oak와 white oak 2수종 간의 구분을 실시한 결과, 68 ~ 84% 정확도로 수종 구분이 가능하다고 보고하였으며, Nisgoski 등(2017)은 동일한 방법론에 의한 4개 수종 구분시 40%의 정확도로 수종 구분이 가능하였다고 보고하였다. 상기와 같은 결과들은 수십 ~ 수백 개 스펙트럼을 이용하여 수종 구분을 실시한 한계들이 존재하였다.

본 장에서는 침엽수 5수종 제재목의 표면에서 측정한 근적외선 스펙트럼을 이용하여 주성분 분석과 SIMCA를 이용한 수종 구분 가능성과 정확도를 평가하고자 하였다.

2. 재료 및 방법

2.1. 공시재료

공시재료는 제 1장에서 밝힌 50 × 100 × 600 mm 크기의 낙엽송(Larch, *Larix kaempferi*), 소나무(Red pine, *Pinus densiflora*), 잣나무(Korean pine, *Pinus koraiensis*), 삼나무(Cedar, *Cryptomeria japonica*), 편백(Cypress, *Chamaecyparis obtusa*) 제재목을 수종별로 50개씩 수집하여 기건한 후 활용하였다.

2.2. 근적외선 스펙트럼 측정과 수학적 전처리

공시재료로부터 근적외선 스펙트럼을 획득하는 방법은 제 1장에서 밝힌 방법과 같으며, 동일한 수학적 전처리 조건에 따라 주성분 분석 및 SIMCA를 실시하였다.

2.3. 주성분 분석을 이용한 군집화 분석

군집화 분석을 위한 주성분 분석은 The Unscrambler version 10.3 (CAMO, Norway)을 이용하여 수행되었다. 주성분 분석에는 근적외선 분광분석기에 의해 수집된 스펙트럼 중 780 nm ~ 2500 nm 대역만을 추출한 스펙트럼을 활용하였다. 수학적 전처리에 의한 수종 구분 정확도를 비교하기 위하여 standard normal variate (SNV), Savitzky-Golay 2^{nd} derivative (derivative option : window size = 21, polynomial order = 2)를 각각 실시하였다. 수학적 전처리에 따라 전체 근적외선 데이터($N_K = 5000$)를 모집단으로 하는 주성분 분석을 실시하였으며, 최대 10개의 주성분 까지 연산하였다. 이에 따라 각 수종별 원 스펙트럼, SNV 전처리를 실시한 스펙트럼, Savitzky-Golay 2^{nd} derivative 전처리를 실시한 스펙트럼에 의한 주성분 분석 결과를 모두 도출하여 군집화를 확인하였다.

2.4. SIMCA를 이용한 수종 구분

SIMCA를 위한 주성분 분석과 SIMCA 수종 구분은 The Unscrambler version 10.3 (CAMO, Norway)을 이용하여 수행하였다. 주성분 분석에는 근적외선 분광분석기를 이용하여 수집된 스펙트럼 중 780 nm ~ 2500 nm 대역만을 추출한 스펙트럼이 활용되었다. 각 수종별 근적외선 데이터($N_K = 1000$)를 8 : 2로 임의 추출하여 800개를 학습 세트(Training set), 200개를 테스트 세트(Test set)로 활용하였다. 주성분 분석에는 학습 세트만이 사용되었으며, 주성분 분석 시 최대 주성분 개수는 10으로 하였다. 이에 따라 각 수종별 주성분 분석 모델을 개발하였다.

SIMCA에서 잔차 평가에 사용되는 각 수종별 주성분 분석 모델의 최적 주성분 개수는 total explained variance를 기준으로 결정하였다. Total explained variance는 원본 데이터의 분산(100%) 대비 주성분 분석 모델이 포함하는 분산을 의미한다. 주성분의 개수가 증가할수록 근적외선 분광분석기의 기계적 잡음 등 무의미한 신호가 포함될 가능성이 높아지기 때문에, 각 수종별 주성분 모델이 포함하는 최적 주성분의 개수는 주성분 증가에 따른 total explained variance의 증가가 1% 미만이 될 때로 결정하였다. 이를 이용하여 SIMCA에 활용될 최적 주성분 모델을 결정하고, 테스트 세트를 이용하여 수종 구분 정확도를 검증하였다.

수학적 전처리에 의한 수종 구분 정확도를 비교하기 위하여 원 스펙트럼에 SNV, Savitzky-Golay 2^{nd} derivative (derivative option : window size = 21, polynomial order = 2)를 각각 실시하였다. 이에 따라 각 수종별 원 스펙트럼, SNV 전처리가 실시된 된 스펙트럼, Savitzky-Golay 2^{nd} derivative 전처리가 실시된 스펙트럼에 의한 주성분 분석 모델을 개발하였다.

SIMCA에서는 학습 세트를 구성하는 데이터들의 잔차 분포와 미지 시험편의 잔차에 의한 유의성 검증을 통해 분류를 실시한다. 본 연구에서는 수종 구분 시, 유의 수준 25%($\alpha = 0.25$)를 기준으로 수종 구분을 실시하였다.

3. 결과 및 고찰

3.1. 수종별 근적외선 스펙트럼 분석

각 수종별 스펙트럼의 평균 흡광도 스펙트럼을 Fig. 2-1 에 나타내었다. 원 흡광도 스펙트럼은 흡광이 넓게 발생하였다. 각 수종별로 스펙트럼의 흡광도 차이가 발생했으며, 그 차이가 파장에 따라 다르게 나타났다.

근적외선은 투과성이 가시광선 등에 비해 떨어지기 때문에, 제재목과 같은 두꺼운 재료는 반사식으로 근적외선 스펙트럼을 측정한다. 이 때, 동일한 재료라 하더라도 재료 표면의 거칠기, 재료의 이방성 등에 의해 입사광의 확산반사율에 차이가 발생한다. 때문에 적절한 수학적 전처리를 수행해야 근적외선 스펙트럼의 분석의 재현성을 확보할 수 있다.

상기와 같은 데이터 처리의 어려움을 해결하기 위해 SNV (Fig. 2-2)와 Savitzky-Golay 2^{nd} derivative (Fig. 2-3)를 실시하였다. SNV 전처리를 실시한 결과, 각 수종간의 평균 스펙트럼이 매우 유사한 흡광도를 나타내게 되었으며, 삼나무의 경우 파장 1100nm 이하의 흡광도가 다른 수종에 비해 높게 나타났다.

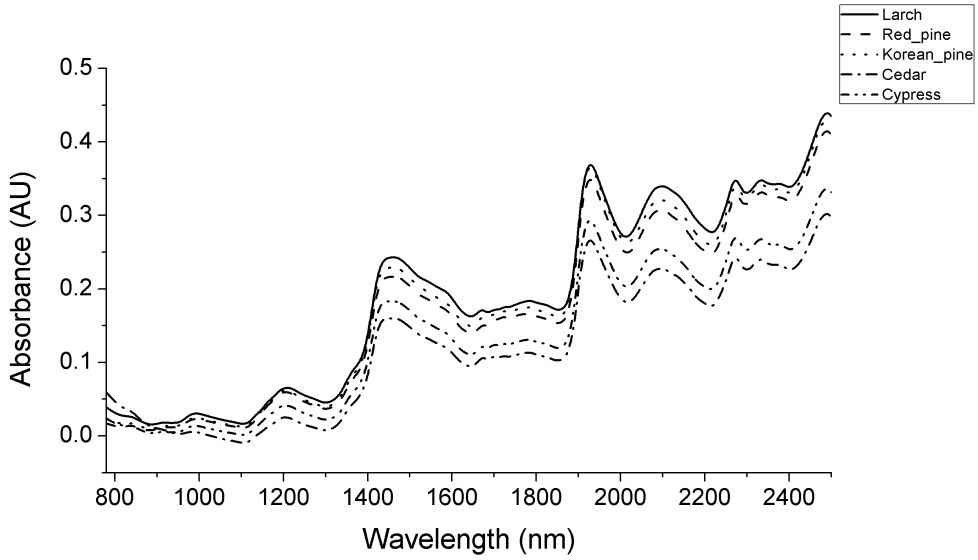


Figure 2-1. Raw average absorbance spectra of each species.

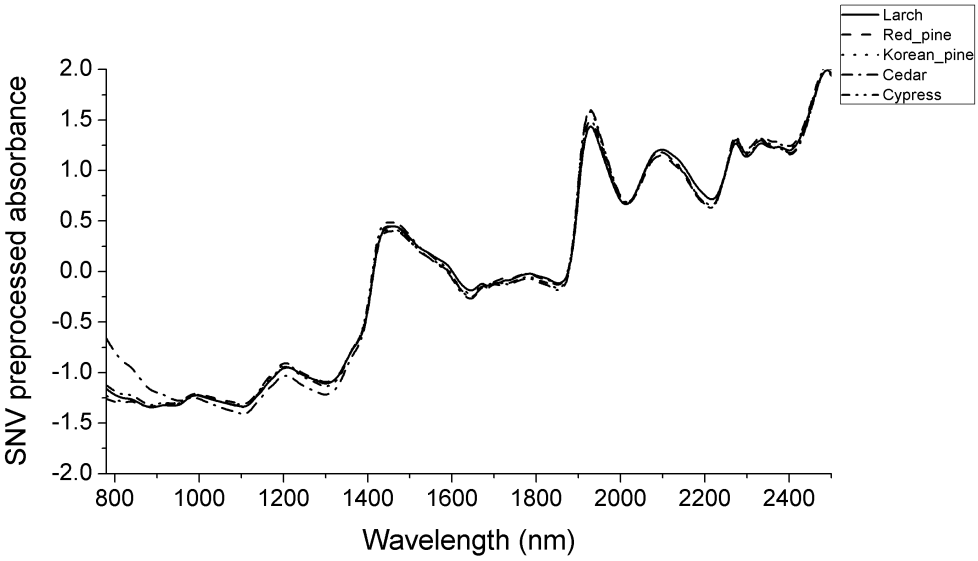


Figure 2-2. Standard normal variate preprocessed average absorbance spectra of each species.

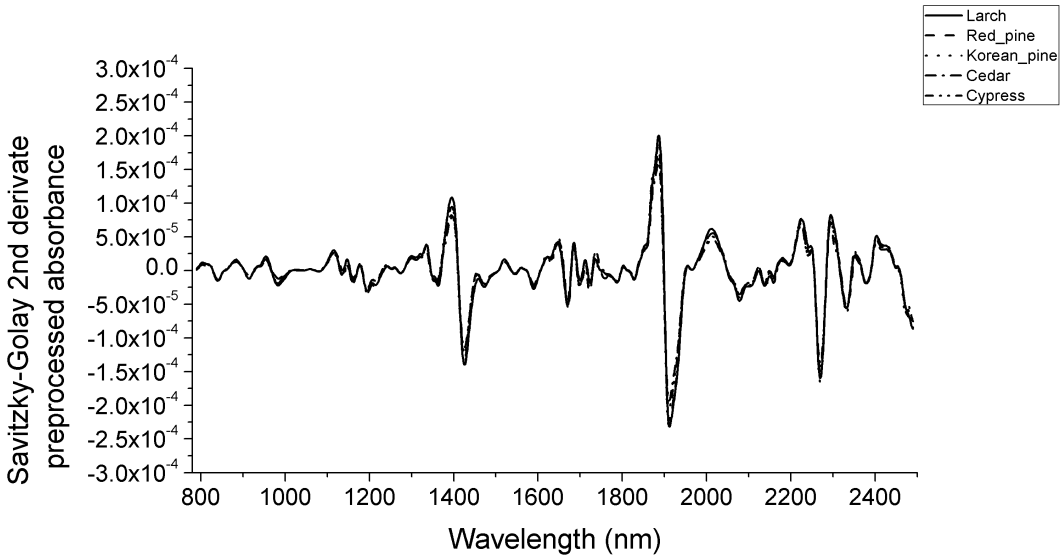


Figure 2-3. Savitzky-Golay 2nd derivative preprocessed average absorbance spectra of each species.

근적외선 흡광지점의 중첩을 해결하기 위한 수단으로 Fig. 2-3과 같이 Savitzky-Golay 2nd derivative 전처리를 수행할 수 있다. Savitzky-Golay 2nd derivative를 수행하였을 때, 근적외선 대역에서 광범위하게 발생하는 흡광의 강한 중첩을 해소함으로써 흡광 데이터의 특징을 부각시킬 수 있다. 전처리를 실시한 스펙트럼의 형태 중 음의 방향으로 나타나는 피크는 중첩된 흡광영역 중 흡광의 중심점, 즉 근적외선을 흡수한 작용기들의 흡광 파장을 의미한다.

Table 2-1은 Fig. 2-4에 나타난 흡광 파장 중 Table Ⅰ-2에 의해 목재의 주요 성분이 갖는 흡광지점으로 밝혀진 부분만을 나타내었다. 기존에 보고된 흡광지점과 본 연구에서 측정한 근적외선 스펙트럼을 이용하여 검출된 흡광지점을 비교한 결과, 각 수종별로 목재 내 주요 화학적 성분인 cellulose, hemicellulose 및 lignin에 존재하는 것으로 알려진 C-H, O-H, C-O 작용기의 진동에 의한 흡광이 발생하는 것을 확인할 수 있었다. Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼의 피크 차이를 직접적으로 비교하기는 어려웠다. 근적외선 스펙트럼을 이용하여 수종 구분을 실시했던 선행 연구들(Horikawa et al., 2015; Hwang 2015)에서도 다른 수종에서 측정한 근적외선 스펙트럼에 Savitzky-Golay 2nd derivative을 실시하더라도 육안상 유사한 패턴을 나타내고 있기 때문에 단순한 특정 위치에서의 스펙트럼의 분석만으로는 수종 구분이 어렵다고 서술한 바 있으며, 이를 해석하기 위해서는 여러 위치에서의 근적외선 스펙트럼의 변화를 추출하는 다변량 분석법이 도입되어야한다고 언급하였다.

Table 2-1. Band assignment of near-infrared absorption occurrence in Fig. 2-4.

Number	Wavelength (nm)	Component	Bond vibration
5	1157	Hemicellulose	2^{nd} OT C-H str.
6	1188-1195	Lignin	2^{nd} OT C-H str.
7	1212-1225	Cellulose	2^{nd} OT C-H str.
8	1350	Hemicellulose	1^{st} OT C-H str. + C-H def.
10	1428	Cellulose/ Water	1^{nd} OT O-H str.
11	1473	Cellulose	1^{st} OT O-H str.
12	1545	Cellulose	1^{st} OT O-H str.
13	1591	Cellulose	1^{st} OT O-H str.
14	1672	Lignin	1^{st} OT C_{ar}-H str.
15	1698	Lignin	1^{st} OT C-H str.
16	1724	Hemicellulose	1^{st} OT C-H str.
18	1790	Cellulose	1^{st} OT C-H str.
20	1916-1942	Water	O-H str. + O-H def. of H_2O
21	2080	Cellulose	O-H str. + C-H def.
24	2270	Cellulose	O-H str. + C-O str.
25	2328-2332	Hemicellulose	C-H str. + C-H def.

Notes) OT : overtone
str. : stretching vibration
def. : deformation vibration

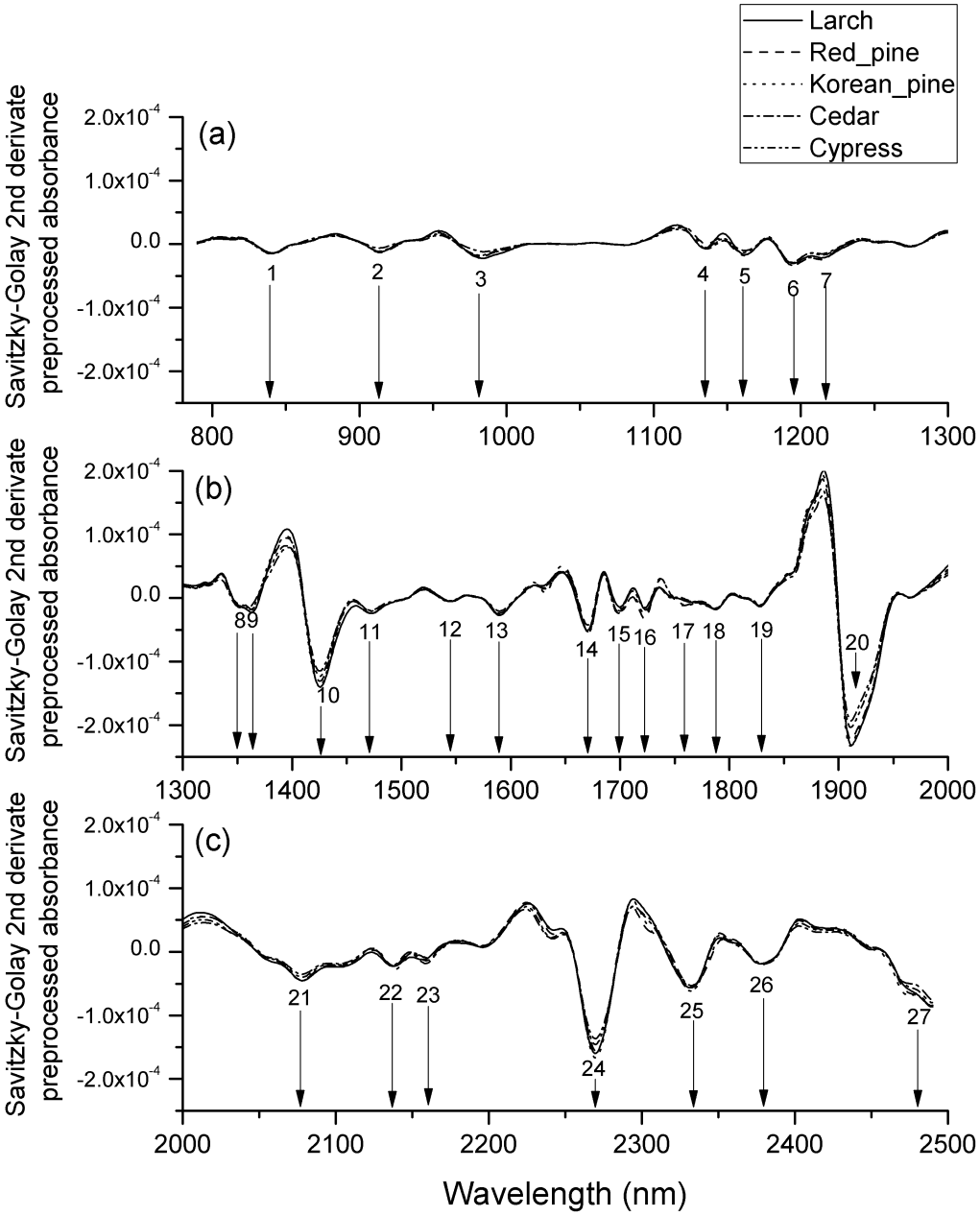


Figure 2-4. Savitzky-Golay 2nd derivative preprocessed absorbance spectra of each species: (a) 780 ~ 1300 nm range, (b) 1300 ~ 2000 nm range, and (c) 2000 ~ 2500 nm range.

3.2. 주성분 분석에 의한 군집화

주성분 분석에 의한 군집화 양상을 분석하기 위하여 본 연구에서 측정한 스펙트럼 전체(N=5000)를 이용하여 수학적 전처리에 따른 주성분 분석을 실시하였다. Total explained variance는 전체 데이터가 갖는 분산 중 첫 r개의 주성분들이 갖는 분산의 비율을 의미한다. 이를 이용하여 주성분들이 갖는 데이터의 정보량을 나타낼 수 있으다. 또한, 데이터를 구성하는 독립변수의 차원 간에 상관도가 높을수록 total explained variance가 높게 나타난다.

Fig. 2-5는 수학적 전처리를 달리한 전체 스펙트럼의 주성분 분석 모델이 갖는 주성분(PC)의 개수에 따른 total explained variance를 나타낸 그래프다. 원 스펙트럼을 이용한 경우 첫 2개의 주성분(PC1 및 PC2)이 약 99%의 total explained variance를 포함하고 있는 것으로 나타났다. 이는 원 스펙트럼의 경우 첫 2개의 주성분을 이용하여 원 데이터의 약 99%까지 복원할 수 있다는 것을 의미한다. 반면 SNV 전처리를 실시한 스펙트럼을 이용한 경우 약 80%, SG 2차 미분 전처리를 실시한 스펙트럼을 이용한 경우 약 70%의 total explained variance를 PC1과 PC2가 포함하고 있었다. SG 2차 미분의 경우 주성분이 증가에 따른 total explained variance가 가장 완만하게 증가하였는데, 이는 미분 효과에 의해 스펙트럼의 중첩 대역이 분리됨에 따라 원 데이터를 구성하는 차원 간 상관성이 감소하였기 때문으로 판단된다.

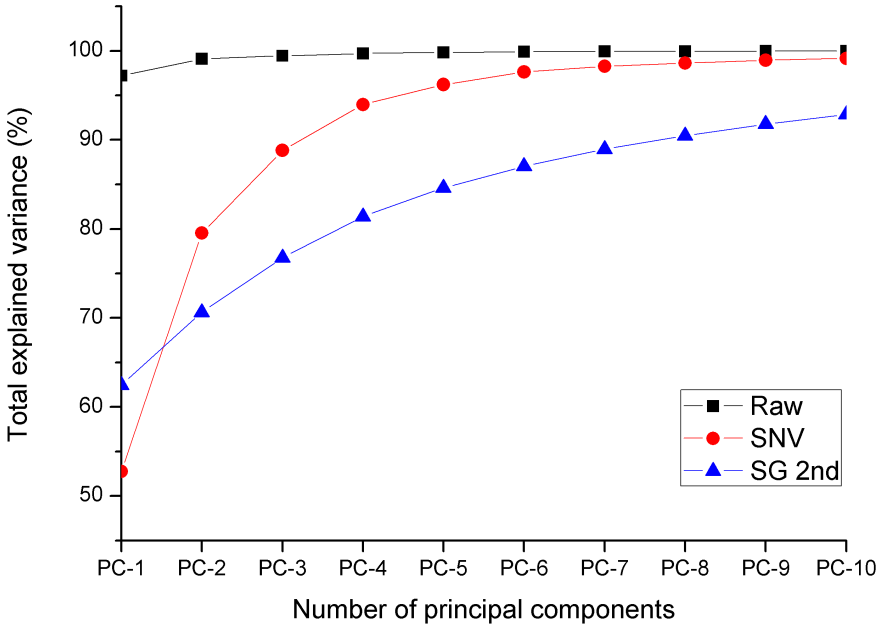


Figure 2-5. Total explained variance of PCA model using raw spectra, standard normal variate (SNV) preprocessed spectra, and Savitzky-Golay 2nd derivative (SG 2nd) preprocessed spectra as a function of the number of principal components

3.2.1. 원 스펙트럼의 군집화 결과

주성분 분석에 의해 결정된 score의 분포를 비교하면 군집화 정도에 따라 수종 구분을 실시할 수 있다. Fig. 2-6은 모든 원 스펙트럼 데이터(N=5000)의 평균 스펙트럼과 PC1과 PC2의 loading을 나타내었으며, Fig. 2-7은 이에 대응하는 PC1과 PC2에서의 score를 나타내었다.

주성분 분석에 의해 추출된 loading과 score를 해석하는 방법은 아래와 같다. 동일한 PC의 loading과 score는 각각 데이터의 평균으로부터 흩어진(잔차의 분산) 방향과 그 정도를 의미하므로, 원 스펙트럼의 주성분 분석 결과(Fig. 2-6)를 토대로 삼나무의 군집화를 해석하면, PC1의 loading은 전 파장대역에서 양의 값을 가지고 있고, 삼나무에서 획득한 스펙트럼들의 PC1 score는 대부분 음의 값을 나타냈으므로, 삼나무에서 측정한 원 흡광도 스펙트럼들은 전체 데이터의 평균보다 전 파장대역에서 흡광도가 낮았다고 해석된다. 또한 스펙트럼의 주성분 분석 결과, PC1의 total explained variance가 97% 였으므로, 평균 흡광도 스펙트럼 값으로부터 PC1의 loading과 각 관측치들의 PC1 score를 곱하여 더해주면 원 데이터의 형태로 평균적으로 97%만큼 복원시킬 수 있다.

원 스펙트럼의 주성분 분석에 의한 PC1-PC2의 score plot을 확인해보면, 낙엽송, 소나무, 잣나무의 score가 1, 3, 4사분면에 걸친 넓은 군집을 형성함을 확인할 수 있었다. 삼나무의 경우 1, 2, 3사분면에 걸친 넓은 군집을 형성하였으며, 낙엽송, 소나무, 잣나무와는 분리되는 것을 확인할 수 있었다. 편백은 양쪽의 군집에 중첩되었다.

위 결과를 종합하면, 전체 수종 원 스펙트럼의 주성분 분석에 의한 PC1-PC2 score 군집 분석으로는 각 수종별 중첩이 강하게 발생하였기 때문에 수종 구분이 불가능할 것으로 판단되었다.

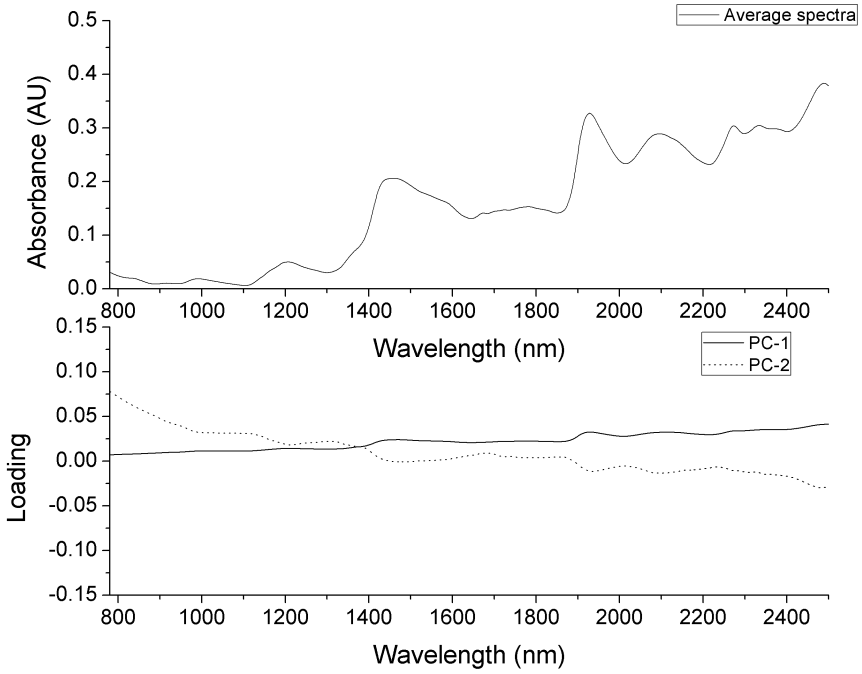


Figure 2-6. Average raw spectra and loading plot (PC1, PC2) of PCA using raw spectra.

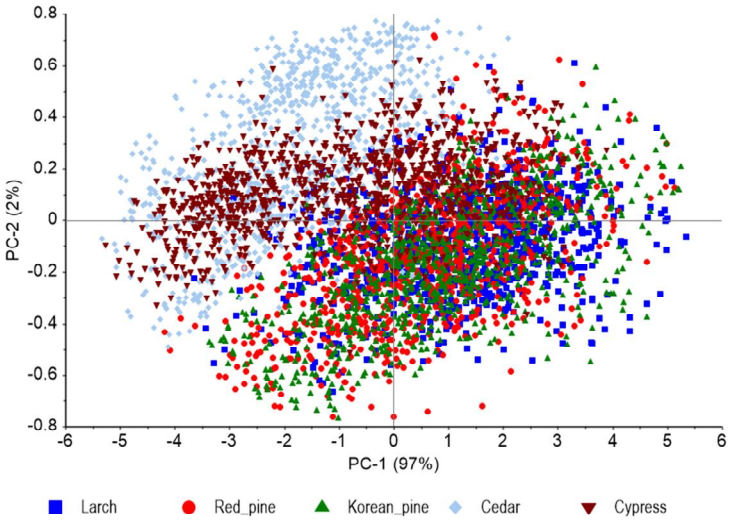


Figure 2-7. Score scatter plot (PC1-PC2) of principal component analysis using raw spectra.

3.2.2. SNV 전처리가 실시된 스펙트럼의 군집화 결과

이러한 현상은 Fig. 2-8 ~ Fig. 2-9에 나타난 SNV 전처리가 실시된 전체 스펙트럼을 이용한 주성분 분석의 결과에서도 관찰되었다. SNV 전처리가 실시된 전체 스펙트럼의 평균 흡광도는 약 -1.5 ~ 2.0 사이의 범위로 변경되었으며, 스펙트럼의 패턴은 원 흡광도 스펙트럼과 큰 차이가 관찰되지 않았다.

Fig. 2-9의 PC1-PC2 score plot을 확인하면, PC1에 의해 삼나무 이외의 모든 수종이 중첩되어 군집을 형성하였다. 낙엽송, 소나무, 잣나무가 강하게 중첩된 군집을 형성하였으며, 편백의 군집은 일부가 낙엽송, 소나무 잣나무와 중첩되었다. 삼나무는 소량 중첩되며 존재하였으나 대부분 PC1에 의해 다른 위치에서 군집을 형성하였다. 따라서 SNV 전처리를 실시한 전체 수종 스펙트럼의 주성분 분석에 의한 PC1-PC2 score 군집 분석으로는 삼나무만이 구분 가능할 것으로 판단되었다. 이는 Fig. 2-8의 PC1 loading과 Fig. 2-2에 나타난 SNV 전처리가 시행된 삼나무의 스펙트럼을 비교해보면, 다른 수종에 비해 높은 흡광도를 나타냈기 때문으로 판단된다.

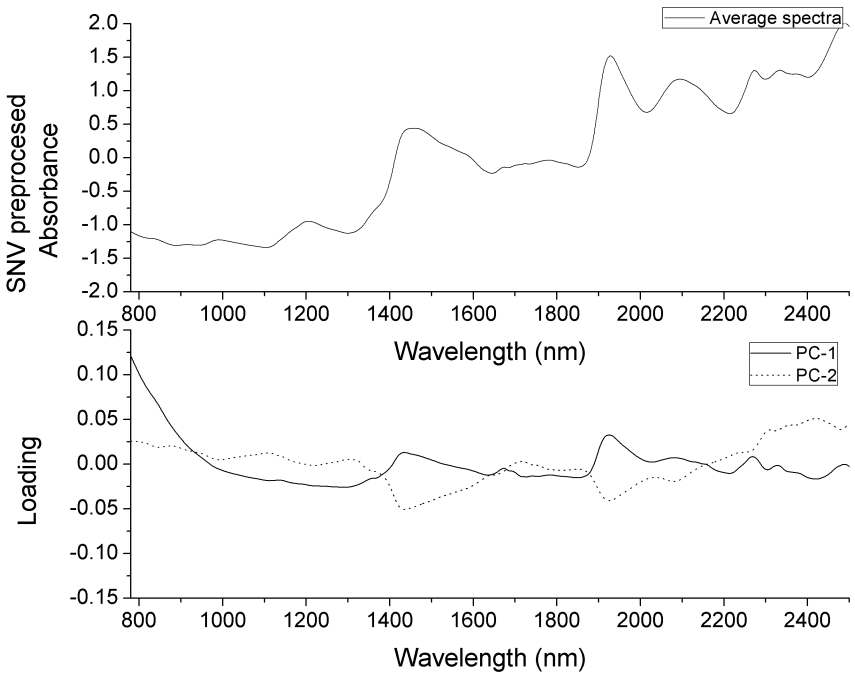


Figure 2-8. Average SNV preprocessed spectra and loading plot (PC1-PC2) of PCA using SNV preprocessed spectra.

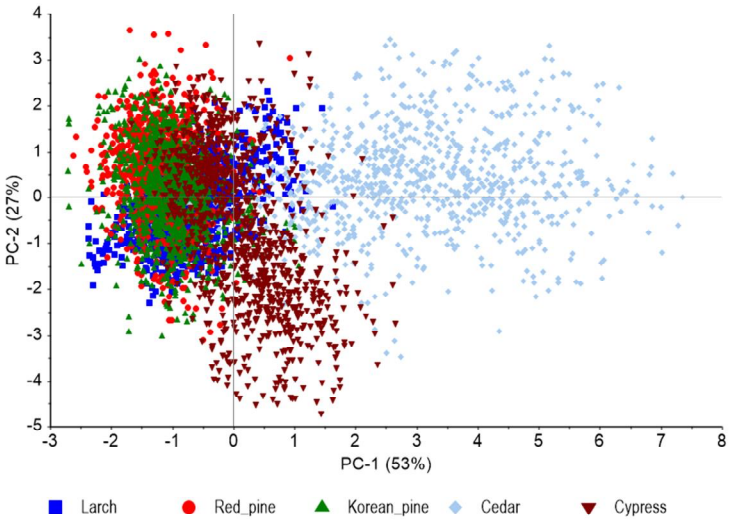


Figure 2-9. Score scatter plot (PC1-PC2) of principal component analysis using SNV preprocessed spectra.

3.2.3. Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼의 군집화 결과

Savitzky-Golay 2nd derivative 전처리가 실시된 스펙트럼을 활용한 주성분 분석의 결과(Fig.2-10 ~ Fig.2-11)에서는 앞선 두 경우와는 다른 군집화 양상을 나타내었다. PC1-PC2 score plot(Fig.2-11)에서는 낙엽송이 하나의 군집, 삼나무와 편백이 하나의 군집, 소나무와 잣나무가 또 다른 하나의 군집을 형성하였다. 이들의 군집은 PC2에 의해 분리되는 것으로 나타났다. 이는 Savitzky-Golay 2nd derivative 전처리가 실시된 근적외선 흡광도 스펙트럼이 낙엽송 / 삼나무, 편백 / 소나무, 잣나무 간에 차이가 발생하고 있다는 것을 의미한다. 그러나 주성분 분석에 의한 score 군집화 형태를 이용해서는 낙엽송 일부, 삼나무와 편백, 잣나무와 소나무 군집으로만 일부 수종 구분이 가능할 것으로 판단되었다.

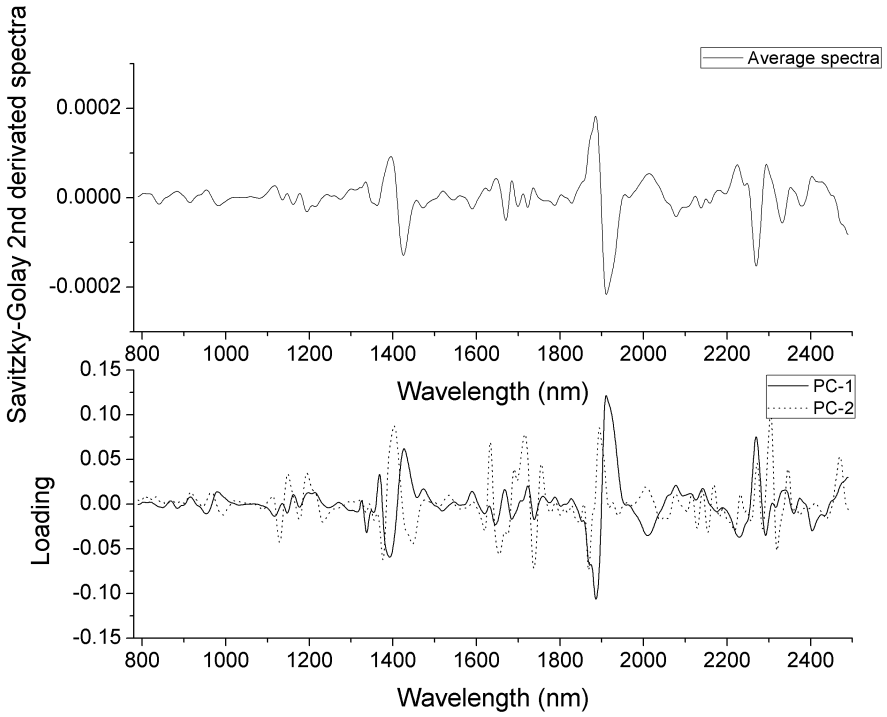


Figure 2-10. Average of Savitzky-Golay 2nd derivative preprocessed spectra and loading plot (PC1-PC2) of PCA using Savitzky-Golay 2nd derivative preprocessed spectra.

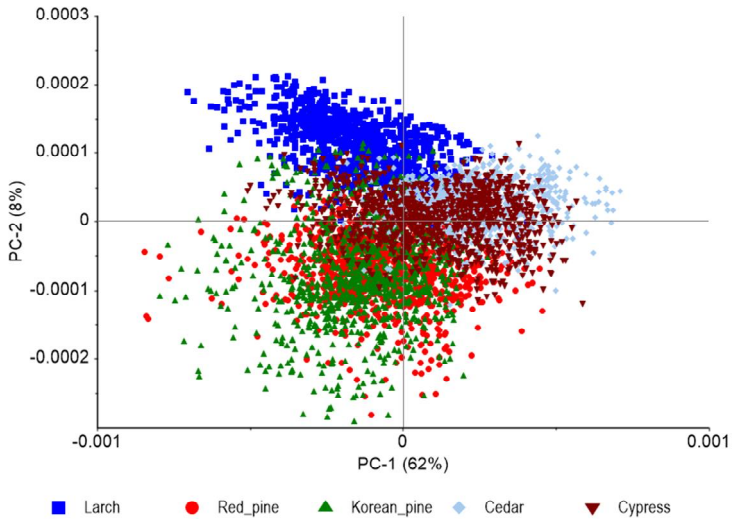


Figure 2-11. PC1-PC2 score plot of PCA using Savitzky-Golay 2nd derivatives preprocessed spectra.

3.3. SIMCA에 의한 수종 구분

3.3.1. 수종별 주성분 분석 모델의 최적 주성분 개수 결정

본 절에서는 주성분 분석에 의한 잔차를 기반으로 수종 구분을 실시하는 SIMCA 실시하기 위하여 각 수종별 근적외선 스펙트럼 중 학습 세트를 이용하여 주성분 분석을 실시하였다. 각 수종별 주성분 분석 결과에 의해 추출된 주성분 중, 잔차 평가에 사용될 최적 주성분의 수는 주성분이 포함하는 데이터의 정보 중 장비에서 유래하는 기계적 잡음 등에 의해 발생하는 과적합을 방지하기 위해 주성분 수 증가에 따른 total explained variance 증가가 1% 미만일 때로 결정하였다.

3.3.1.1 원 스펙트럼을 이용한 주성분 분석 모델

원 흡광도 스펙트럼 중 임의 선발된 학습 세트(각 수종별 800개 스펙트럼)를 이용하여 SIMCA에 사용될 주성분 모델을 개발하였다. 원 스펙트럼의 경우 스펙트럼의 중첩이 높아 흡광도 사이의 상관도가 높고 기준선 보정 등이 수행되지 않은 상태이기 때문에, Fig. 2-5에서 나타난 바와 같이 전체 원 흡광도 스펙트럼을 이용한 주성분 분석 결과에서는 PC1에 의해 97.22%의 분산이 포함되었다. 이러한 경향은 각 수종별 스펙트럼을 이용한 결과에서도 마찬가지로 나타났으며, Fig. 2-12에 나타난 바와 같이 낙엽송의 주성분 모델은 2개(99.33%), 소나무의 주성분 모델은 2개(99.10%), 잣나무 주성분 모델은 2개(99.28%), 삼나무 주성분 모델은 2개(98.71%), 및 편백 주성분 모델은 1개(98.98%)의 주성분을 사용할 때 최적인 것으로 결정되었다.

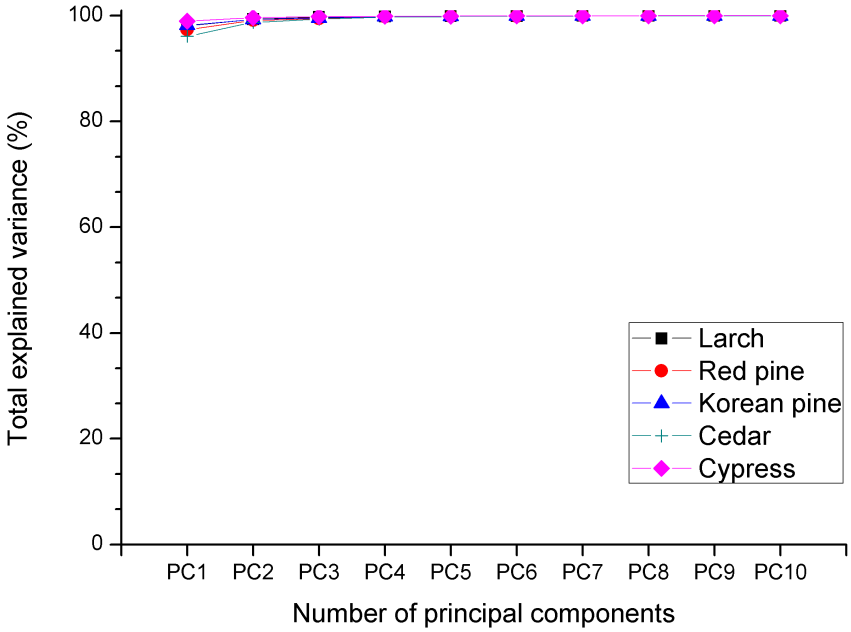


Figure 2-12. Total explained variance of PCA model using each species raw spectra as a function of the number of principal components.

3.3.1.2 SNV 전처리를 실시한 스펙트럼을 이용한 주성분 분석 모델

SNV 전처리를 실시한 흡광도 스펙트럼 중 임의 선발된 학습 세트(각 수종별 800개 스펙트럼)를 이용하여 SIMCA에 사용될 주성분 모델을 개발하였다. SNV 전처리를 실시한 경우, 기준선 보정과 산란 보정 효과에 의해 Fig. 2-2에 나타난 바와 같이 각 수종별 스펙트럼 사이의 위치 변이가 감소하였다. 이에 따라 넓은 흡광영역을 가지더라도 스펙트럼간의 상관성이 약화됨에 따라 Fig. 2-5에서 나타난 바와 같이 전체 SNV 전처리를 실시한 흡광도 스펙트럼을 이용한 주성분 분석 결과에서 PC1에 의해 52.75%의 분산만이 포함되었다. 이러한 경향은 각 수종별 스펙트럼을 이용한 결과에서도 마찬가지로 나타났다. SIMCA에 사용될 각 수종별 SNV 전처리를 실시한 스펙트럼을 이용한 주성분 모델의 최적 주성분의 개수를 total explained variance(Fig. 2-13)를 이용하여 선정한 결과, 낙엽송의 주성분 모델은 7개(97.55%), 소나무의 주성분 모델은 7개(96.73%), 잣나무 주성분 모델은 6개(96.05%), 삼나무 주성분 모델은 5개(97.00%), 및 편백 주성분 모델은 6개(98.32%)의 주성분을 사용할 때 최적인 것으로 결정되었다.

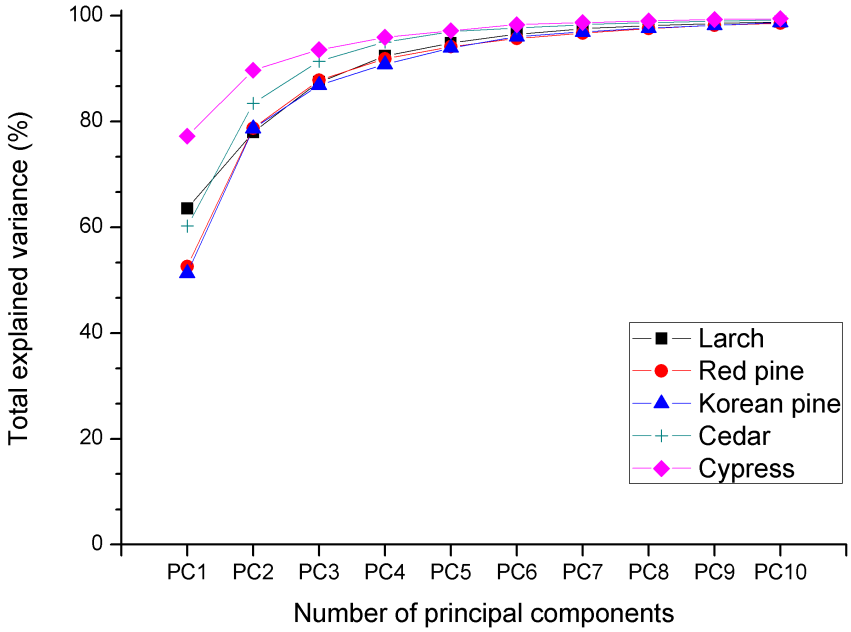


Figure 2-13. Total explained variance of PCA model using each species standard normal variate preprocessed spectra as a function of the number of principal components.

3.3.1.3 Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼을 이용한 주성분 분석 모델

Savitzky-Golay 2nd derivative 전처리를 실시한 흡광도 스펙트럼 중 임의 선발된 학습 세트(각 수종별 800개 스펙트럼)를 이용하여 SIMCA에 사용될 주성분 모델을 개발하였다. Savitzky-Golay 2nd derivative 전처리를 실시한 경우, 기준선 보정과 흡광 영역 분리에 의해 Fig. 2-3에 나타난 바와 같이 각 수종별 스펙트럼 사이의 위치 변이가 감소하였으며, 흡광도가 부각되었다. 이에 따라 Fig. 2-5에서 나타난 바와 같이 전체 Savitzky-Golay 2nd derivative 전처리를 실시한 흡광도 스펙트럼을 이용한 주성분 분석 결과에서 PC1에 의해 62.47%의 분산만이 포함되었다. 이러한 경향은 각 수종별 스펙트럼을 이용한 결과에서도 마찬가지로 나타났다. SIMCA에 사용될 각 수종별 Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼을 이용한 주성분 모델의 최적 주성분의 개수를 total explained variance(Fig. 2-14)를 이용하여 선정한 결과, 낙엽송의 주성분 모델은 7개(91.30%), 소나무의 주성분 모델은 9개(91.11%), 잣나무 주성분 모델은 9개(91.93%), 삼나무 주성분 모델은 8개(89.54%), 및 편백 주성분 모델은 6개(91.11%)의 주성분을 사용할 때 최적인 것으로 결정되었다.

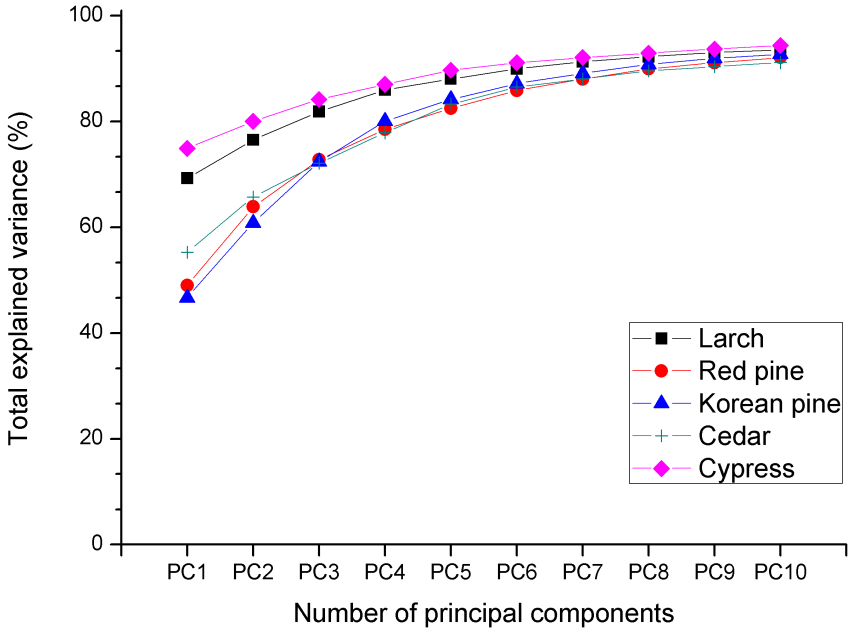


Figure 2-14. Total explained variance of PCA model using each species Savitzky-Golay 2nd derivative preprocessed spectra as a function of the number of principal components.

3.3.2. 원 스펙트럼을 이용한 수종 구분 결과

원 스펙트럼을 이용한 각 수종별 최적 주성분 모델을 토대로 SIMCA에 의한 수종 구분을 실시한 결과를 Table 2-2에 나타내었다. 분류 모델의 신뢰도를 평가에는 정확도, 정밀도 및 재현율을 활용한다. 정확도(Accuracy)는 평가에 활용된 전체 데이터의 수 중 정확하게 분류된 비율을 의미한다. 정밀도(Precision)은 분류 모델에 의해 집단 K로 분류된 데이터 중 실제 집단 K인 데이터의 수를 의미한다. 재현율(Recall)은 실제 집단 K 인 데이터의 수 중 집단 K로 예측된 데이터의 비율을 의미한다. 원 스펙트럼을 이용한 SIMCA 수종 구분 정확도는 35.5%로 나타났다. 정밀도는 수종에 따라 99.07% ~ 78.95% (낙엽송, 삼나무, 편백, 잣나무, 소나무 순)로 나타났다. 그러나 재현율은 수종에 따라 57.5% ~ 16% (삼나무, 낙엽송, 편백, 잣나무, 소나무 순)으로 매우 낮았다.

전체적으로 미분류(Unassigned)와 중복 분류(Multi-classified)가 다량 발생하였다. 중복 분류 중 2 중복 분류, 3 중복 분류, 4 중복 분류 및 5 중복 분류도 일부 존재하였다. 특히 소나무와 잣나무간의 중복 분류가 높게 나타났다. 편백 또한 삼나무와의 중복 분류의 비중이 높은 발생한 반면, 삼나무의 경우 편백과의 중복 분류의 비중이 낮았다.

Table 2-2. Classification results of SIMCA based on each species PCA models using raw spectra.

		Actual Species					
		Larch	Red pine	Korean pine	Cedar	Cypress	Precision (%)
Predicted species	Larch	106	0	1	0	0	99.07
	Red pine	0	30	6	0	2	78.95
	Korean pine	1	3	32	0	1	86.49
	Cedar	1	1	0	115	2	96.64
	Cypress	2	1	2	5	72	87.80
Unassigned		52	48	36	60	29	
Multi-classified	2 species	19	82	78	13	30	
	3 species	13	32	38	4	41	
	4 species	6	3	7	2	16	
	5 species	0	0	0	1	7	
Total		200	200	200	200	200	
Recall (%)		53.00	15.00	16.00	57.50	36.00	
Accuracy (%)		35.50					

3.3.3. SNV 전처리를 실시한 스펙트럼을 이용한 수종 구분 결과

SNV 전처리를 실시한 스펙트럼을 이용한 각 수종별 최적 주성분 모델을 토대로 SIMCA에 의한 수종 구분을 실시한 결과를 Table 2-3에 나타내었다. SNV 전처리를 실시한 스펙트럼을 이용한 SIMCA 수종 구분 정확도는 51.9%로 원 스펙트럼에 기반한 SIMCA 분류에 비해 개선된 성능을 나타내었다. SNV 전처리를 실시한 스펙트럼을 이용한 SIMCA 수종 구분의 정밀도는 수종에 따라 100% ~ 90.67%(낙엽송, 편백, 소나무, 잣나무, 삼나무 순)로 나타났으며, 모두 원 스펙트럼의 결과에 비해 개선되었다. 그러나 재현율은 수종에 따라 76.5% ~ 19.5%(낙엽송, 삼나무, 잣나무, 소나무, 편백 순)로 나타났으며, 편백의 재현율을 제외하고는 모두 원 스펙트럼에 기반한 주성분 모델 분류에 비해 개선된 결과를 나타내었다.

원 스펙트럼을 이용한 SIMCA 수종 구분에 비해 미분류된 시험편의 개수는 소나무와 삼나무를 제외하고 감소하였으며, 3, 4, 5 중복 분류가 현저하게 감소하였음을 확인할 수 있었다. 편백의 경우 삼나무와의 중복분류가 대부분이었고, 잣나무와 소나무간의 중복분류가 여전히 약 20 ~ 30% 가량 존재하였다.

Table 2-3. Classification results of SIMCA based on each species PCA models using standard normal variate preprocessed spectra.

		Actual Species					
		Larch	Red pine	Korean pine	Cedar	Cypress	Precision (%)
Predicted species	Larch	153	0	0	0	0	100.00
	Red pine	0	94	3	0	0	96.91
	Korean pine	0	5	97	0	0	95.10
	Cedar	0	1	0	136	13	90.67
	Cypress	0	0	0	0	39	100.00
Unassigned		43	56	43	64	27	
Multi-classified	2 species	4	40	57	0	112	
	3 species	0	4	0	0	9	
	4 species	0	0	0	0	0	
	5 species	0	0	0	0	0	
Total		200	200	200	200	200	
Recall (%)		76.50	47.00	48.50	68.00	19.50	
Accuracy (%)		51.90					

3.3.4. Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼을 이용한 수종 구분 결과

Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼을 이용한 주성분 분석 결과를 토대로 SIMCA 수종 구분을 실시한 결과를 Table 2-4에 나타내었다. SG 2차 미분 전처리를 실시한 스펙트럼을 이용한 SIMCA 수종 구분 정확도는 73%로, 앞선 두 경우에 비해 가장 정확도가 가장 높았다. 정밀도는 수종에 따라 100% ~ 98.54%(낙엽송, 삼나무, 편백(이상 100%), 소나무, 잣나무 순)로 평가되었다. 그러나 재현율은 수종에 따라 82.5% ~ 67.5%(삼나무, 낙엽송, 소나무, 편백, 잣나무 순)로 평가되었으며, 낙엽송을 제외하고 SNV 전처리를 실시한 스펙트럼을 활용한 분류에 비해 개선된 결과를 나타내었다.

원 스펙트럼을 이용한 SIMCA 수종 구분 결과에 비해 미분류된 시험편의 개수는 낙엽송과 삼나무에서만 감소하였으나, 중복 분류가 현저하게 감소하였음을 확인할 수 있었다. 편백의 경우 편백과 삼나무간의 중복 분류가 대부분을 차지하였으며, 잣나무와 소나무는 상호간의 중복 분류가 여전히 존재하였다. 그러나 앞선 두 결과에 비해 중복 분류에 의한 오류가 상당부분 해소되었음을 확인할 수 있었다.

Table 2-4. Classification results of SIMCA based on each species PCA models using Savitzky-Golay 2nd derivative preprocessed spectra.

		Actual Species					
		Larch	Red pine	Korean pine	Cedar	Cypress	Precision (%)
Predicted species	Larch	152	0	0	0	0	100.00
	Red pine	0	139	1	0	0	99.29
	Korean pine	2	0	135	0	0	98.54
	Cedar	0	0	0	165	0	100.00
	Cypress	0	0	0	0	139	100.00
Unassigned		37	52	41	33	32	
Multi-classified	2 species	9	9	23	2	23	
	3 species	0	0	0	0	5	
	4 species	0	0	0	0	1	
	5 species	0	0	0	0	0	
Total		200	200	200	200	200	
Recall (%)		76.00	69.50	67.50	82.50	69.50	
Accuracy (%)		73.00					

4. 결론

본 장에서는 상관도가 높은 다차원 근적외선 스펙트럼을 이용한 수종별 분류를 실시하기 위하여 주성분 분석을 이용한 score 군집화 분석과 이를 기반으로 한 SIMCA 수종 구분을 실시하였다.

공시 재료로부터 확보한 모든 원 흡광도 스펙트럼, SNV 전처리를 실시한 흡광도 스펙트럼 및 Savitzky-Golay 2^{nd} derivative 전처리를 실시한 흡광도 스펙트럼을 이용하여 각각의 주성분 분석을 통해 score를 이용한 군집 분석을 시도한 결과, 각 수종별 데이터의 score 중첩으로 인하여 군집 분석으로는 수종 구분이 곤란하였다.

각 수종별 데이터를 주성분 분석하여 추출한 주성분 모델을 이용하여 SIMCA 분류를 실시한 결과, 원 스펙트럼을 이용한 SIMCA 분류의 정확도는 35.5%, 최소 정밀도는 78.95%, 최소 재현율은 15%였으며, 소나무의 분류 성능이 가장 떨어졌고, 중복 분류가 빈번하게 발생하였다. SNV 전처리를 실시한 스펙트럼을 이용하여 SIMCA 분류를 실시하였을 때의 정확도는 51.9%, 최소 정밀도는 90.67%, 최소 재현율은 19.5%로 나타났으며, 원 스펙트럼을 이용한 결과에 비해 분류 성능이 향상되고 중복 분류의 비율이 현저히 감소하였다. Savitzky-Golay 2^{nd} derivative 전처리를 실시한 스펙트럼을 이용한 SIMCA 분류 결과, 정확도는 73%, 최소 정밀도는 98.54%, 재현율은 67.5%로 다른 두 경우에 비해 가장 개선된 성능을 나타내었다. 그러나, 정확도 수준이 낮았다.

따라서 제재목의 표면에서 측정한 근적외선 스펙트럼을 이용하여 SIMCA에 의해 수종 구분을 실시하기 위해서는 가장 정밀도가 높았던 Savitzky-Golay 2^{nd} derivative 전처리를 실시한 모델을 활용하는 것이 가장 적합할 것으로 판단되나, 낮은 정확도와 재현율 제고를 위한 개선방법 개발이 요구되었다.

제 3장

부분 최소 자승 판별 분석을 이용한 수종 구분

1. 서론

근적외선 분광분석법에서 활용되는 부분 최소 자승법은 목재의 화학적 주성분 함량비 분석(Hodge et al. 2004; Alves et al. 2011; Silva et al. 2013; He et al. 2013) 및 물성 예측(Tsuchikawa 2007; Via 2010; Fujimoto et al. 2012; Tyson et al; 2012; Eom et al. 2013, Yang et al., 2017)에서 강력한 성능을 나타내고 있다. 부분 최소 자승법을 이용한 목재 수종 구분에 대한 연구는 Antti 등(1996)에 의해 처음 시도되었는데, 3수종의 목재 칩을 일정 비율로 섞어 분말화한 후 근적외선 스펙트럼을 이용한 목재 분말 내 각 수종의 질량비를 예측함으로써 수종 구분에 대한 가능성을 제시하였다. 근적외선 스펙트럼을 이용한 수종 구분에 관한 연구는 브라질에서 활발하게 연구 중에 있다. Pastore 등(2011)과 Braga 등(2011)은 마호가니 유사 수종을 구분하기 위하여 각각 목분/소시편에서 획득한 근적외선 스펙트럼을 이용한 부분 최소 자승 판별 분석을 수행하였다. Pastore는 마호가니와 육안상 유사한 4개 수종의 구분을 위하여 목분 수십 점에서 측정한 근적외선 스펙트럼을 이용하여 부분 최소 자승 판별 분석을 실시한 결과, 성공적으로 수종 구분이 가능하였다고 보고하였다. 이후 수종을 확장시키는 연구(Soares et al., 2017)와 동일 수종의 생산지 차이(국가)에 따른 구분 가능성에 관한 연구 결과(Bergo et al., 2016)를 보고하였다.

제재목의 단면에서 측정한 스펙트럼은 목분에 비해 재현성이 낮아 수종 구분 정확도가 떨어진다는 연구 결과(Nisgoski et al., 2015)가 보고된 바 있으나, 현장 적용성 및 활용도를 제고하기 위하여 제재목에서 측정한 근적외선 스펙트럼으로 수종 구분을 실시하는 것이 바람직하다(Ramalho et al., 2018). 목분에서 측정한 근적외선 스펙트럼을 이용한 부분 최소 자승 판별 분석 결과(Pastore et al., 2011; Bergo et al., 2016), 100%에 가까운 정확도로 수종 구분이 가능하였다. 그러나, 제재목을 활용하는 연구에서는 연구에 사용된 수종과 시험편/스펙트럼의 수에 따라 구분 정확도에 차이가 나타났다(Braga et al., 2011 – 100%; Lazarescu et al.,

2017 - 89.8%; Ramalho et al., 2018 - 70%). 이 외에도 Horikawa 등(2015)과 Hwang 등(2015)은 해부학적으로 유사한 소나무 2수종의 구분을 부분 최소 자승 판별 분석을 통해 수행한 바 있다.

본 장에서는 국산 침엽수 5수종의 수종 구분을 위하여 근적외선 스펙트럼을 이용한 부분 최소 자승 판별분석에 의한 수종 구분 정확도를 평가하고자 하였다.

2. 재료 및 방법

2.1. 공시재료

공시재료는 제 1장에서 밝힌 50 × 100 × 600 mm 크기의 낙엽송(Larch, *Larix kaempferi*), 소나무(Red pine, *Pinus densiflora*), 잣나무(Korean pine, *Pinus koraiensis*), 삼나무(Cedar, *Cryptomeria japonica*), 편백(Cypress, *Chamaecyparis obtusa*) 제재목을 수종별로 50개씩 수집하여 기건한 후 활용하였다.

2.2. 근적외선 스펙트럼 측정과 수학적 전처리

공시재료로부터 근적외선 스펙트럼을 획득하는 방법은 제 1장에서 밝힌 방법과 같으며, 동일한 수학적 전처리 조건에 따라 부분 최소자승 판별 분석을 실시하였다.

2.3. 부분 최소 자승 판별 분석 모델 개발을 위한 종속변수 설정

부분 최소 자승 판별 분석은 독립변수인 스펙트럼에 종속변수로 class value를 지정하여 선형 회귀 분석하는 분류법이다. 본 연구는 5종의 침엽수종에 대하여 구분하는 것을 목적으로 하므로, 다음 Table 3-1과 같은 5개의 class value를 갖도록 하는 부분 최소 자승 판별 분석을 실시하였다.

Table 3-1. Model structure for PLS-DA.

	Species				
	Larch	Cedar	Korean pine	Red pine	Cypress
	Class value				
Model Larch	1	0	0	0	0
Model Cedar	0	1	0	0	0
Model Korean pine	0	0	1	0	0
Model Red pine	0	0	0	1	0
Model Cypress	0	0	0	0	1

2.4. 부분 최소 자승 판별 분석 모델 개발

부분 최소 자승 판별 분석은 PLS-TOOLBOX ver 8.6.2 (Eigenvector research incorporated, US)를 활용하였다. 부분 최소 자승 판별 분석 모델 개발 시, 판별 분석 모델의 신뢰성 검증을 위해 k-fold 교차 검정(k-fold cross validation, Fig. 3-1)을 실시하였다. k-fold 교차 검정은 전체 학습 데이터를 k개의 집단으로 분리하고, k-1개의 집단으로 부분 최소 자승 판별 모델을 학습 세트로 하여 모델을 개발한 후, 모델 개발에 사용되지 않았던 1개의 집단으로 모델의 성능을 평가하는 것을 모든 집단이 평가 대상이 될 때까지 반복함으로써 모델의 신뢰성을 확보하는 방법이다. 따라서 교차 검정 과정에서 산출되는 오차를 이용해 최적 모델의 과적합을 방지하고 회귀 모델의 범용성을 확보할 수 있다. 실제 모델 개발에는 데이터의 균등한 분배와 재현성 확보를 위하여 venetian blind에 의한 10-fold 교차 검정이 실시되었다. 부분 최소 자승법에 의한 모델 개발 시, 최대 요인 수(Number of latent variables)는 20으로 설정하였다.

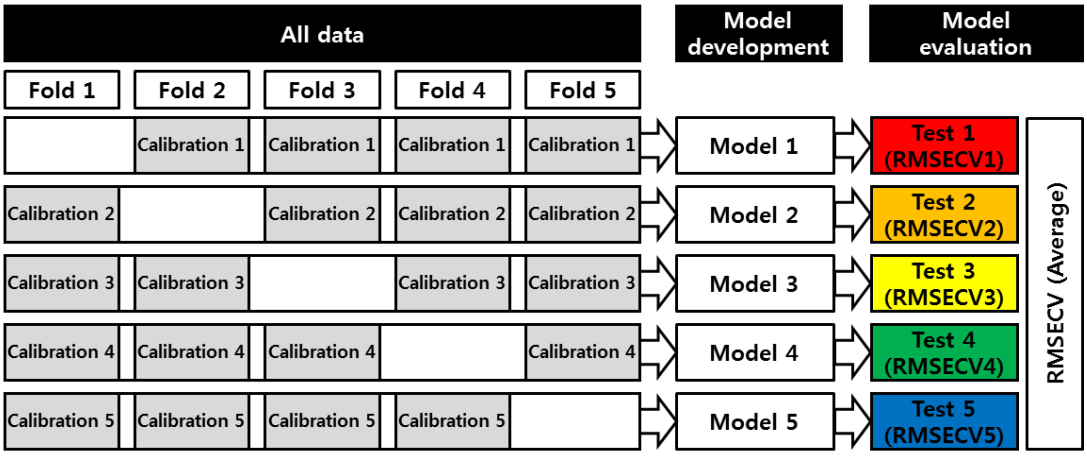


Figure 3-1. Schematic diagram of k-fold cross validation (k=5).

2.5. 부분 최소 자승 판별 분석 모델의 최적 요인 수 결정

부분 최소 자승법을 이용한 모델 개발 시, 회귀 모델이 갖는 요인 수가 증가함에 따라 학습 세트의 예측 성능은 증가하게 된다. 회귀 모델이 갖는 요인 수를 높임으로써 학습 세트의 예측 성능을 증가시킬 수는 있으나, 회귀 모델의 학습 세트에 대한 예측 과적합(Overfitting)을 유발하게 된다. 이는 모델 개발에 사용된 데이터에는 적합하지만, 학습에 사용되지 않은 외부 데이터에 대한 예측 성능을 감소시키는 요인이 된다. 따라서 회귀 모델의 과적합을 방지하기 위하여 반드시 검증 과정에 의해 적절한 요인 수를 결정해야한다. 본 연구에서는 부분 최소 자승법에 의한 회귀 모델의 최적 요인 수 선정에 교차 검정에 의한 예측 정확도를 활용하였다.

부분 최소 자승법을 포함한 대부분의 회귀 모델의 성능은 관측치의 종속변수 측정값 y_i과 모델에 의한 예측된 종속변수 예측값 $y_{i,predicted}$ 사이의 오차를 전체 데이터($i=1, \cdots, N$)에 대해 평가함으로써 아래의 Eq. Ⅲ-1로 표현되는 예측 오차 제곱합(Predicted error sum of squares, PRESS) 또는 Eq. Ⅲ-2로 표현되는 평균 제곱근 오차(Root mean squared error, RMSE)로서 평가한다.

$$PRESS = \sum_{i=1}^{N}(y_i - y_{i,predicted})^2 \qquad \text{(Eq. Ⅲ-1)}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i - y_{i,predicted})^2}{N}} \qquad \text{(Eq. Ⅲ-2)}$$

교차 검정을 실시한 경우, 일반적으로 Fig. 3-2(Poon et al. 2012)와 같이 요인 수의 증가에 따라 학습 세트(검량 세트, calibration set)의 평균 제곱근 오차(Root mean squared error of calibration, RMSEC)는 지속적으로 감소하는 것에 비해 교차 검정 데이터의 평균 제곱근

오차(Root mean squared error of cross validation, RMSECV)는 증가한다. 이는 학습에 활용되지 않은 교차 검정 세트가 갖는 외부 자료로서의 효과 때문이다. 따라서 학습 데이터에 대한 과적합을 방지하고 회귀 모델의 외부 데이터에 대한 예측 신뢰도를 확보하기 위하여 부분 최소 자승법에 의한 회귀 모델의 최적 요인 수는 RMSECV가 최소값을 나타낸 때로 정한다.

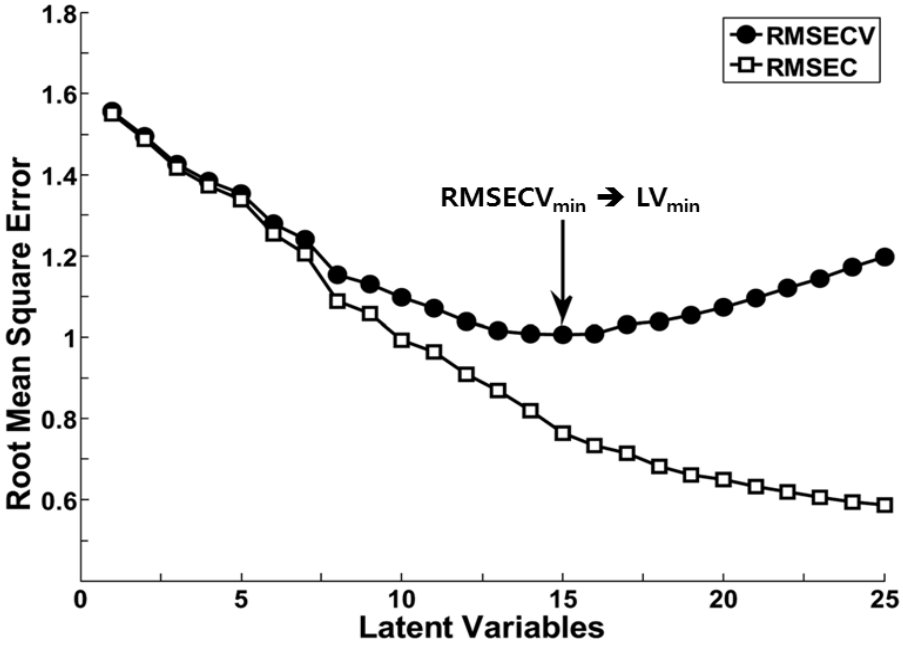


Figure 3-2. An example plot of the root mean square error of calibration (RMSEC) versus root mean square error of cross validation as a function of the number of latent variables. The optimum number of latent variables corresponds to the minimum of RMSECV (Poon et al. 2012)

그러나, 상기와 같은 기준은 요인 수 증가에 따른 PRESSCV (Predicted error sum of squares of cross validation) 또는 RMSECV의 최소값이 존재하지 않고 지속적으로 감소하는 경우에는 적용이 어렵다. 이를 위해 Wold(1978)는 요인 수 증가에 의한 모델의 성능 개선율을 반영하는 R criteron(Eq. Ⅲ-3)을 정의하여 최적 요인 수를 결정하는 방법을 제안하였다.

$$R = PRESSCV(l+1)/PRESSCV(l) \quad (l = number\,of\,latent\,variables) \qquad \text{(Eq. Ⅲ-3)}$$

Wold가 최초 제안한 R criterion의 개념은 l개의 요인을 포함하는 모델의 교차 검정 예측 오차 ($PRESSCV(l)$) 대비 $l+1$개의 요인을 포함하는 모델의 교차 검정 예측 오차($PRESSCV(l+1)$) 가 1 이상이 되는 첫 요인 수 l를 찾아 이를 최적 요인 수로 결정하는 방법이다. 이는 상술한 최소 RMSECV를 찾는 과정과 동일하다. 그러나 교차 검정에서의 예측 오차가 지속적으로 감소하는 경우(R값이 항상 1 이하인 경우)에 대한 적용 불가능성, R값이 1일 때 항상 외부 데이터에 대한 적합성을 보장하지 못한다는(Krzanowski 1987, Osten 1988) 문제 등에 의해 현재에는 1 미만의 조정된 R criterion (0.9 ~ 0.95)을 사용하여 최적 요인 수를 결정하는 것이 일반적(Li et al., 2002)이다. 조정된 R 값이 의미하는 바는 개발자가 정하는 유의미한 성능 개선의 한계를 정하는 것이다.

부분 최소 자승 판별 분석에서는 물성이 아닌 집단에 따라 1 또는 0이 되는 모조 종속 변수(Dummy dependent variable)를 갖도록 하는 모델을 개발한다. 예를 들어, 낙엽송의 부분 최소 자승 판별 분석 모델의 경우 낙엽송에서 획득한 스펙트럼의 예측치를 1, 이외 수종의 스펙트럼 예측치는 0으로 만드는 부분 최소 자승 모델을 개발하는 것이다. 부분 최소 자승 판별 분석에서의 RMSECV는 정의에 따라 모조 종속 변수에 대한 편차를 나타낸다.

분류를 목적으로 할 때, RMSECV는 목표 하는 class value로의 평균적인 예측 밀집도를 나타내기는 하지만 모델에 의한 두 집단의 예측 값이 0 또는 1에서 편향(bias)된 분포를 나타내는 경우, 분류에는 문제가 없더라도 RMSECV는 증가한다. 이 때문에 부분 최소 자승 판별 분석에서는 RMSECV가 분류 정확성을 지표하지 못한다. 그러나, 부분 최소 자승 판별 분석을 실시한 많은 선행 연구에서 RMSECV와 R criterion에 의한 모델 선정을 하고 있다(Braga et al., 2011; Almeida et al., 2013; Soares et al., 2017). 정확한 집단 구분 성능을 요인 수 선정을 위해 RMSECV 대신 교차 검정 분류 오류율(Classification error ratio of cross validation, e_{cv})을 활용하는 것이 최적 요인 수 결정에 더 적합하며(Kjeldahl and Rasmus, 2010). 본 연구에서는 이를 이용하여 최적 요인 수를 선정하였다.

집단 K 판별 분석 모델이 갖는 교차 검정 오류율($e_{CV,K}$)은 다음과 같이 정의된다.

$$e_{CV,K} = \frac{\frac{n(y_{predicted,\,CV,K} < y_{criteria})}{N_K} + \frac{n(y_{predicted,\,CV,not\,K} \geq y_{criteria})}{N_{not\,K}}}{2}$$

(Eq. Ⅲ-4)

교차 검정 오류율은 실제로는 집단 K 이지만, 교차 검정에 의한 예측치($y_{predicted,\,CV,K}$)가 판별 경계치($y_{criteria}$) 미만으로 예측되어 집단 $not\,K$ 로 잘못 판별된 비율($n(y_{predicted,\,CV,K} < y_{criteria})/N_K$)과 실제로 집단 $not\,K$ 지만 교차 검정에 의한 예측치($y_{predicted,\,CV,not\,K}$)가 판별 경계치 이상으로 예측되어 집단 K로 판별된 비율($n(y_{predicted,\,CV,not\,K} \geq y_{criteria})/N_{not\,K}$)간의 평균을 의미한다.

K 개의 집단 간의 다중 분류 문제에서 판별 분석 모델의 평균 교차 검정 분류 오류율(E_{CV})은 Eq. Ⅲ-5와 같이 정의된다.

$$E_{CV} = \sum_{i=1}^{K} \frac{e_{CV,i}}{K} \qquad \text{(Eq. Ⅲ-5)}$$

본 연구에서는 수종 판별 분석 모델의 최적 요인수 선정을 위해 Wold가 제안한 R criterion과 평균 교차 검정 세트의 분류 오류율 E_{CV} 를 결합하여, 다음 Eq. Ⅲ-6과 같이 E^* criterion을 정의하였으며, E^*가 0.95 이상이 되는 첫 번째 l을 수종 판별 분석 모델의 최적 요인 수로 결정하였다.

$$E^* = E_{CV}(l+1)/E_{CV}(l) \qquad \text{(Eq. Ⅲ-6)}$$

2.6. 부분 최소 자승 판별 분석 모델을 이용한 다중 집단의 수종 판별

부분 최소 자승 판별 분석은 1 또는 0이 되는 모조 종속 변수(Dummy dependent variable)를 갖도록 하는 모델을 개발한다. 일반적으로 종속변수를 이진화(0 또는 1)하여 생성하므로 판별 기준치를 0.5로 설정한다. 이는 수종 판별 분석 모델의 예측치가 0.5 이상이면 해당 수종, 0.5 미만인 경우에는 해당 수종이 아닌 것으로 판정한다는 것을 의미한다. 독립 변수와 종속 변수간의 선형 상관관계를 모델화하는 부분 최소 자승법은, 비선형적 상관관계를 갖는 근적외선 스펙트럼과 1 또는 0이 되는 모조 종속 변수간의 완벽한 모델링을 보장하지 못한다. 때문에 집단 K 판별 모델의 예측치가 1 또는 0를 기준으로 분포하지 않고 편향되거나(biased) 넓게 분포하는 문제가 발생할 수 있다. 이를 해결하기 위하여 1장의 3.2.4.절에 밝힌 판별 모델의 예측치의 확률변환에 의한 판별 기준을 적용하였다.

두 개 집단 사이의 분류를 실시하는 2진 분류(Binary classification)에서는 예측치를 이용하여 기준치에 따른 판별을 수행하지만, 다중 집단 분류(Multi-class classification)의 경우 각 집단 판별 모델에 의해 출력된 복수의 예측치로 소속 집단 판별을 수행해야 한다. 이에 따라 본 연구에서는 동일한 각 수종의 판별 분석 모델의 $y_{predicted,CV}$ 값에 대하여, 2가지 수종 판별 방법에 따른 수종 분류 성능을 비교하였다.

(1) $y_{predicted} \geq 0.5$ 인 경우가 1개의 수종 판별 분석 모델에서만 존재하는 경우에만 해당 수종으로 분류하며, 이외의 경우는 기각한다.

(2) $F_{norm}(y_{predicted}INK) \geq 0.5$ 인 경우가 1개의 수종 판별 분석 모델에서만 존재하는 경우에 해당 수종으로 분류하며, 이외의 경우는 기각한다.

3. 결과 및 고찰

3.1. 부분 최소 자승 판별 분석 모델의 최적 요인 수 결정

3.1.1. 원 스펙트럼을 이용한 부분 최소 자승 판별 분석 모델

원 스펙트럼을 이용하여 개발한 각 수종별 부분 최소 자승 판별 분석 모델의 요인 수에 따른 각 수종별 교차 검정 분류 오류율($e_{CV,K}$)을 Fig. 3-3에 나타내었다. 각 수종별 판별 분석 모델의 $e_{CV,K}$ 는 수종에 따른 차이는 있으나 부분 최소 자승 판별 분석 모델이 갖는 요인 수 증가에 의해 계단상으로 감소하는 것으로 나타났다.

원 스펙트럼을 이용한 부분 최소 자승 판별 분석 모델의 최적 요인 수를 선정하기 위해 Eq. Ⅲ-5를 이용하여 교차 검정 평균 분류 오류율(E_{CV})을 평가한 결과, 요인 수 증가에 따라 평균 분류 오류율은 지속적으로 감소하였다. E_{CV} 를 이용하여 Eq. Ⅲ-6에 의해 E^*를 평가하였을 때, E^*가 0.95를 초과하는 첫 번째 요인의 수는 8개($E^* = 0.96$)일 때로 나타났으며(Fig. 3-4), 이 때의 각 수종별 판별 분석 모델의 교차 검정 평균 분류 오류율(E_{CV})은 7.76%로 나타났다. 이에 따라 원 스펙트럼을 이용한 부분 최소 자승 판별 분석 모델의 최적 요인 수는 8개로 결정되었다.

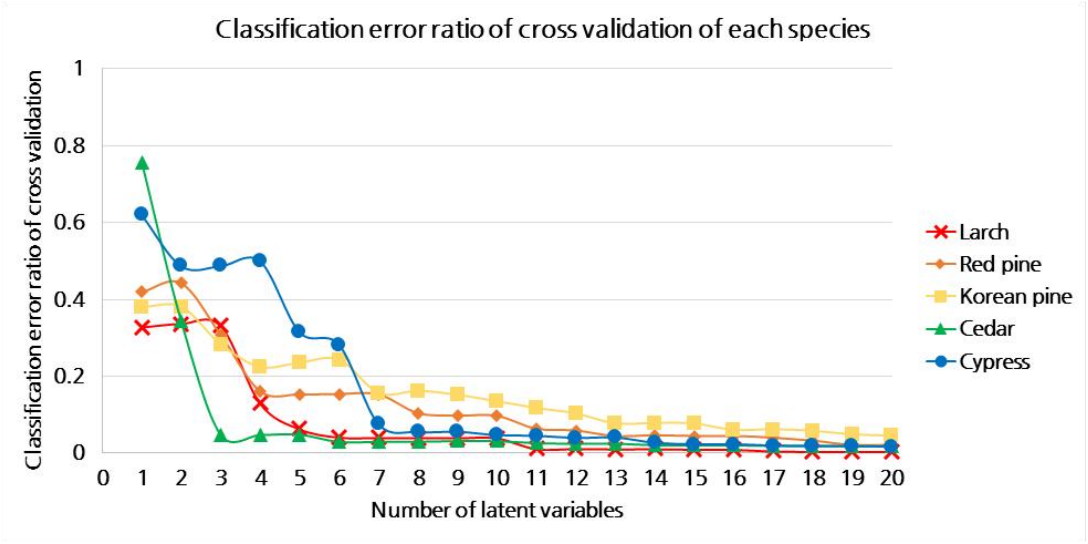


Figure 3-3. Classification error ratio of cross validation as a function of the number of latent variables for each species partial least squares discriminant analysis using raw spectra.

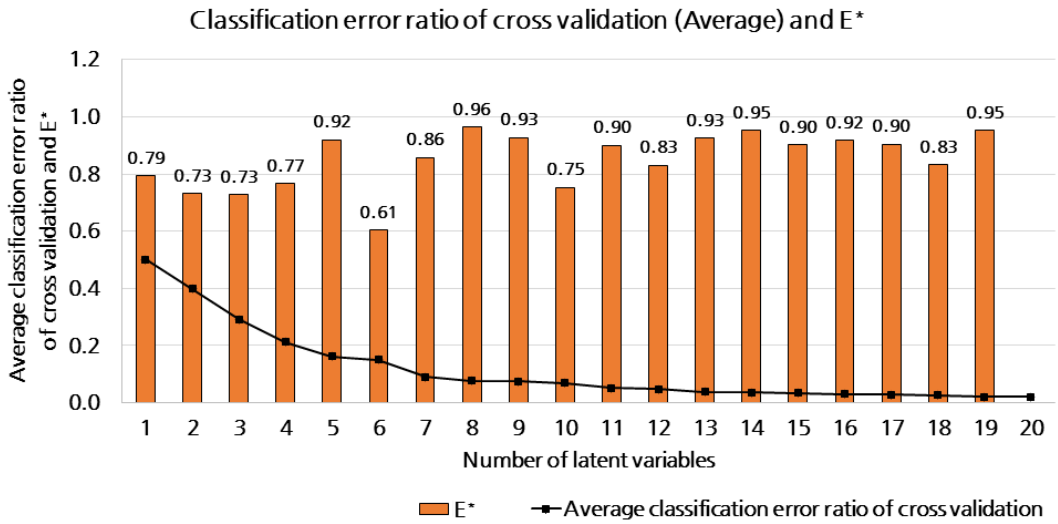


Figure 3-4. Average classification error ratio of cross validation and E* values as a function of the number of latent variables for each species partial least squares discriminant analysis model using raw spectra.

Fig. 3-5는 원 스펙트럼을 이용한 각 수종별 부분 최소 자승 판별 분석 모델에서 나타난 교차 검정 예측치($y_{predicted,\,CV}$)의 분포를 나타내었다. 모든 수종별 부분 최소 자승 판별 모델에 의한 각 집단들의 $y_{predicted,\,CV}$ 는 정규분포로 근사 가능한 것으로 판단되었다.

부분 최소 자승 판별 분석 모델은 Table 3-1과 같이 판별을 원하는 수종을 1, 이외의 수종을 0으로 하는 모델을 개발하는 것임에도 불구하고, 모든 수종별 판별 분석 모델들이 갖는 해당 수종 집단의 교차 검정 예측치($y_{predicted,\,CV,K}$)는 평균은 0과 1사이에 존재하였으며 그 외 집단($y_{predicted,\,CV,not\,K}$)의 분포 또한 마찬가지였다. 특히 소나무와 잣나무 판별 분석 모델의 경우 해당 수종 집단의 교차 검정 예측치($y_{predicted,\,CV,K}$)의 평균이 0.5 인근으로 나타나 원 스펙트럼을 이용한 수종 구분은 정확도가 낮을 것으로 판단되었다.

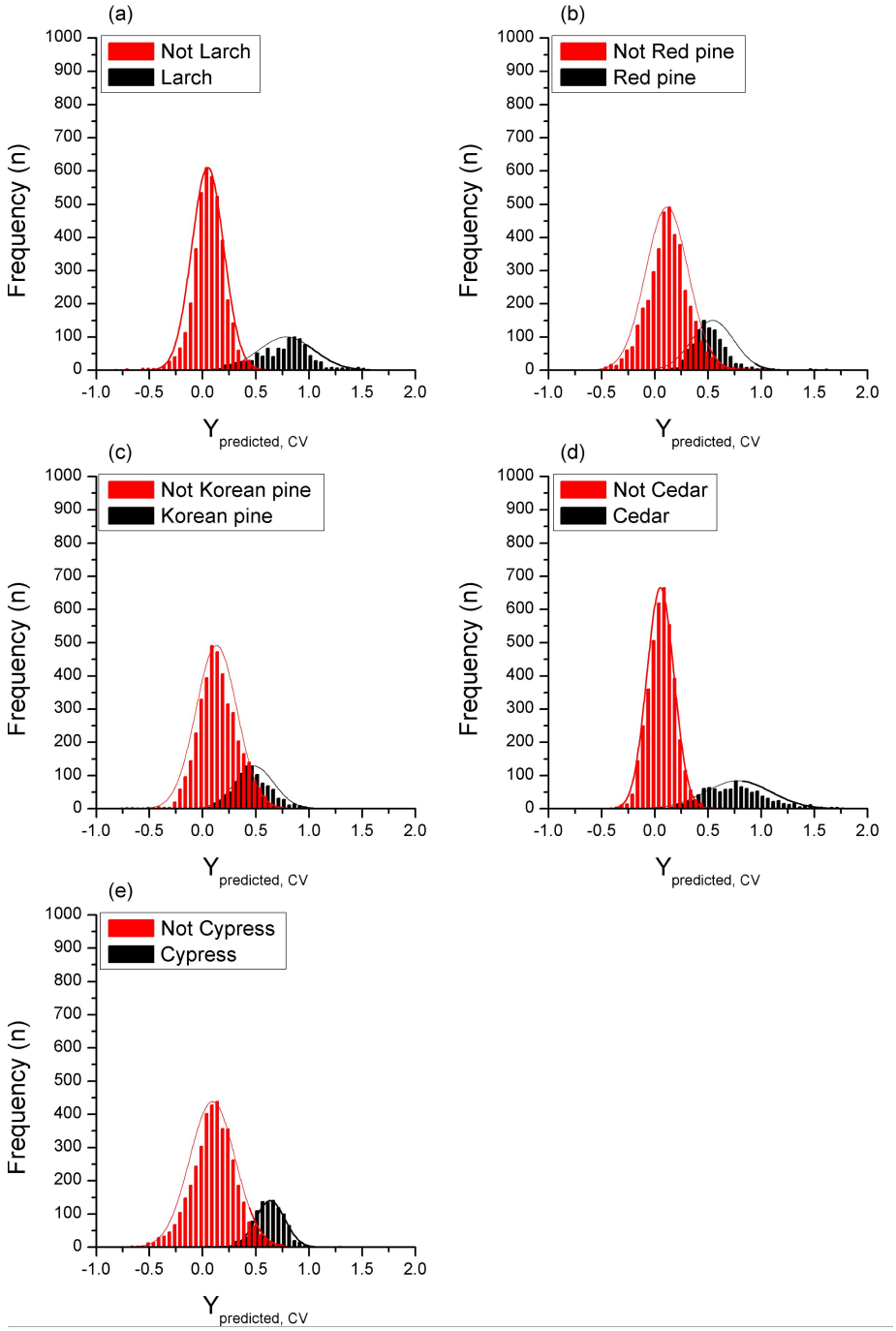


Figure 3-5. Histogram of predicted Y of cross validation by partial least squares discriminant analysis model with 8 latent variables using raw spectra; (a) model Larch, (b) model Red pine, (c) model Korean pine, (d) model Cedar and (e) model Cypress

3.1.2. SNV 전처리를 실시한 스펙트럼을 이용한 부분 최소 자승 판별 분석 모델

SNV 전처리를 실시한 스펙트럼을 이용하여 개발한 각 수종별 부분 최소 자승 판별 분석 모델의 요인 수에 따른 각 수종별 교차 검정 세트의 분류 오류율($e_{CV,K}$)을 Fig. 3-6에 나타내었다. 각 수종별 판별 분석 모델의 $e_{CV,K}$ 는 수종에 따른 차이는 있으나 요인 수 증가에 의해 계단상으로 감소하는 것으로 나타났으며, 원 스펙트럼을 이용한 결과에 비해 빠르게 감소하였다.

SNV 전처리를 실시한 스펙트럼을 이용한 부분 최소 자승 판별 분석 모델의 최적 요인 수를 선정하기 위해 교차 검정 세트의 평균 분류 오류율(E_{CV})을 평가한 결과, 요인 수 증가에 따라 평균 분류 오류율은 지속적으로 감소하였다. E_{CV}를 이용하여 E^*를 평가한 결과(Fig. 3-7), E^* 값이 0.95를 초과하는 첫 번째 요인 수는 7개($E^*=1.00$)로 나타났으며 이 때의 E_{CV} 는 7.98%로 나타났다.

이에 따라, SNV 전처리를 실시한 스펙트럼을 이용한 부분 최소 자승 판별 분석 모델의 최적 요인 수는 7개로 결정되었다.

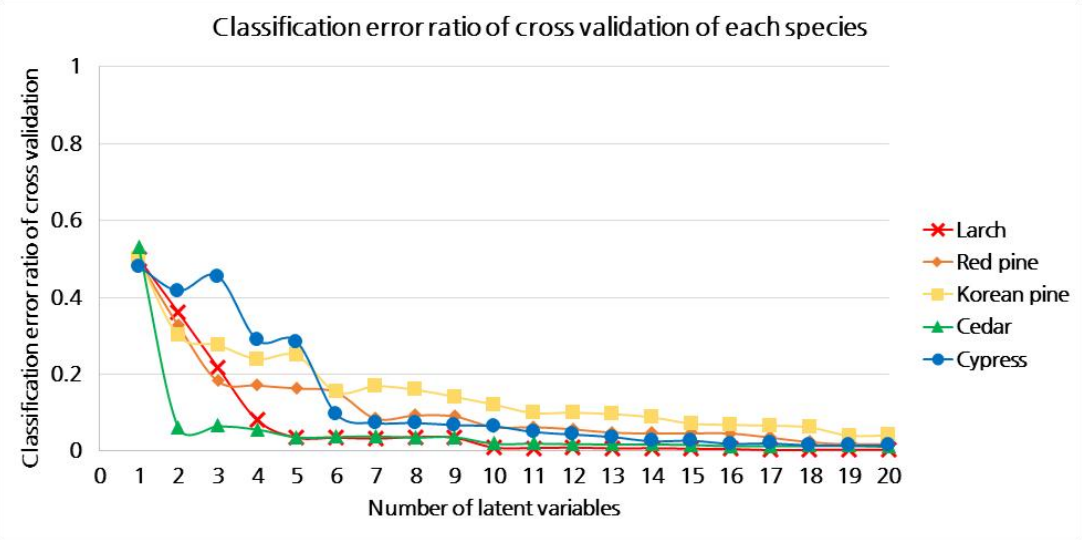


Figure 3-6. Classification error ratio of cross validation as a function of the number of latent variables for each species partial least squares discriminant analysis model using SNV preprocessed spectra.

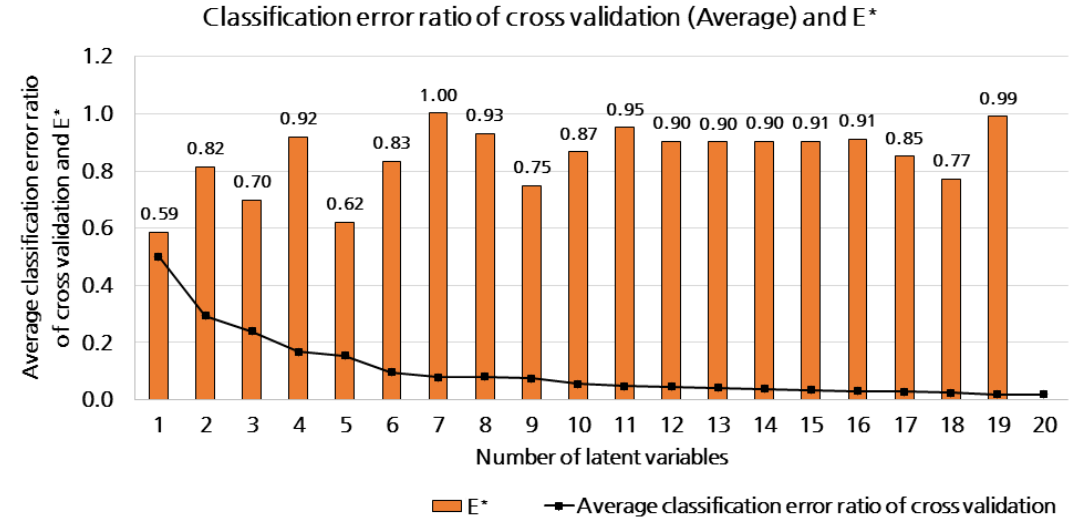


Figure 3-7. Average classification error ratio of cross validation and E* values as a function of the number of latent variables for each species partial least squares discriminant analysis model using SNV preprocessed spectra.

Fig. 3-8는 SNV 전처리를 실시한 스펙트럼을 이용한 각 수종별 부분 최소 자승 판별 분석 모델의 교차 검정 예측치($y_{predicted,\,CV}$)의 분포를 나타내었다. 잣나무, 삼나무, 편백의 부분 최소 자승 판별 모델에 의한 각 집단들의 $y_{predicted,\,CV}$ 는 정규분포로 근사 가능한 것으로 판단되었으나 낙엽송과 잣나무는 정규분포를 벗어나는 것으로 판단되었다.

모든 수종별 판별 분석 모델들이 갖는 해당 수종 집단의 예측치($y_{predicted,\,CV,K}$)는 원 스펙트럼을 이용한 결과와 동일하게 0과 1사이의 평균을 가졌으며, 그 외 집단($y_{predicted,\,CV,not\,K}$)의 분포 또한 마찬가지였다. 특히 소나무와 잣나무 판별 분석 모델의 경우 $y_{predicted,\,CV,K}$ 의 평균이 0.5 인근으로 나타나 집단간 구분이 원활히 이루어지지 않았다. 따라서, SNV 전처리는 수종 구분은 원 스펙트럼을 이용한 경우에 비해 구분 성능 개선 효과가 미미할 것으로 판단되었다.

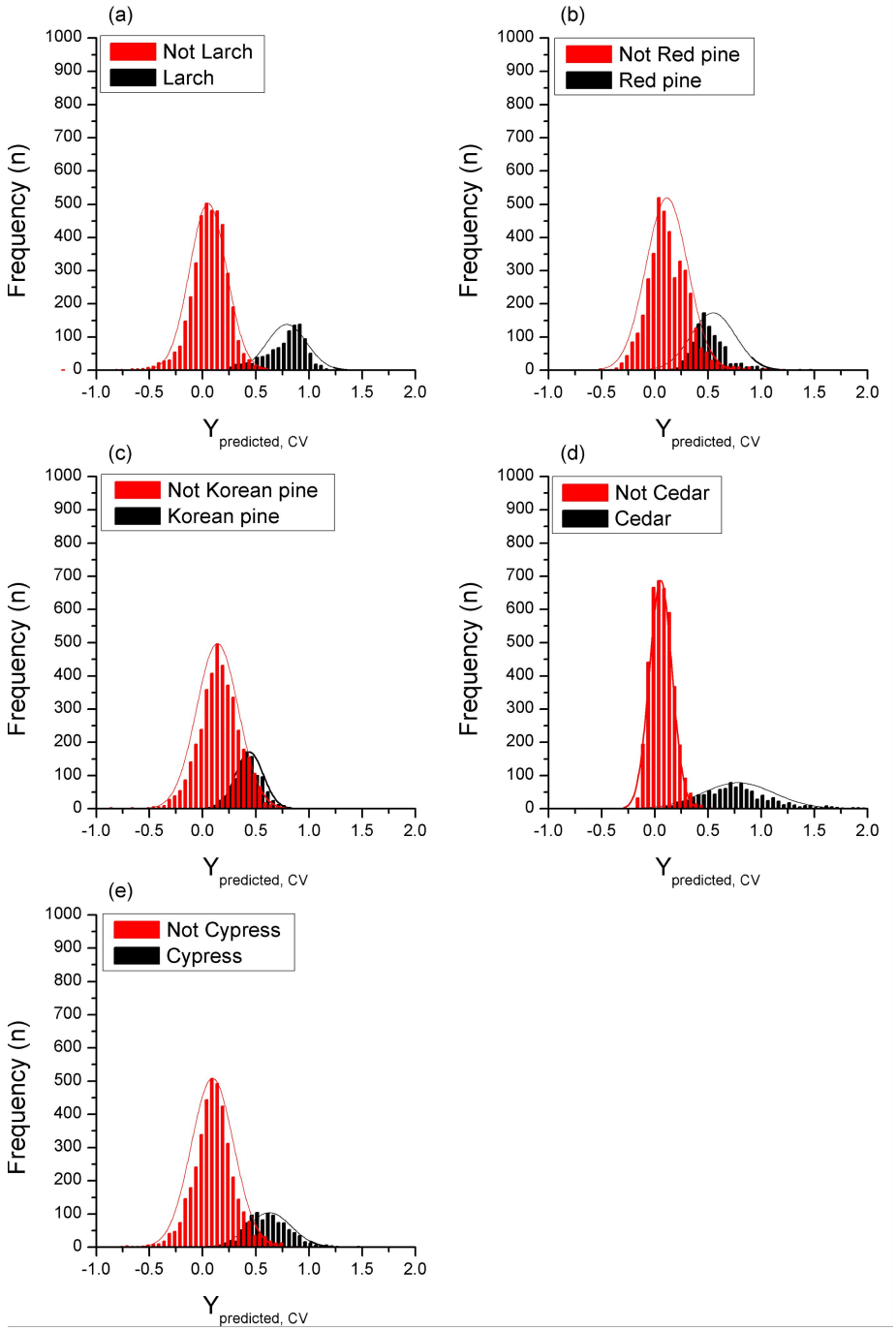


Figure 3-8. Histogram of predicted Y of cross validation by partial least squares discriminant analysis model with 7 latent variables using standard normal variate preprocessed spectra; (a) model Larch, (b) model Red pine, (c) model Korean pine, (d) model Cedar and (e) model Cypress.

3.1.3. Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼을 이용한 부분 최소 자승 판별 분석 모델

Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼을 이용하여 개발한 각 수종별 부분 최소 자승 판별 분석 모델의 요인 수에 따른 각 수종별 교차 검정의 분류 오류율($e_{CV,K}$)을 Fig. 3-9에 나타내었다. 각 수종별 판별 분석 모델의 $e_{CV,K}$ 는 앞선 두 경우에 비해 요인 수 증가에 의해 초기 단계에서 급격하게 감소하는 것으로 나타났다.

Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼을 이용한 부분 최소 자승 판별 분석 모델의 최적 요인 수를 선정하기 위해 교차 검정 평균 분류 오류율(E_{CV})을 평가한 결과, 요인 수 증가에 따라 평균 분류 오류율은 지속적으로 감소하는 경향을 나타냈다. E_{CV} 를 이용하여 E^*를 평가한 결과(Fig. 3-10), E^* 값이 0.95를 초과하는 첫 번째 요인 수는 14로 나타났으며 ($E^*=1.05$), 이 때의 각 수종별 판별 분석 모델의 교차 검정 평균 분류 오류율(E_{CV})은 0.78%로 원 스펙트럼을 이용한 모델 및 SNV 전처리를 실시한 스펙트럼을 이용한 모델의 결과에 비해 1/10 수준으로 나타났다.

이에 따라 Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼을 이용한 부분 최소 자승 판별 분석 모델의 최적 요인 수는 14개로 결정되었다.

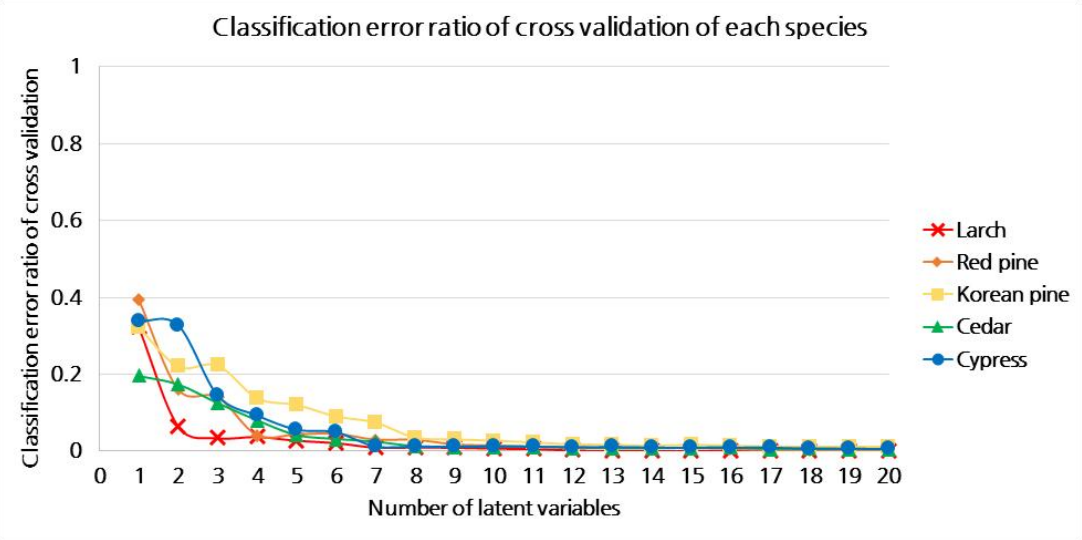


Figure 3-9. Classification error ratio of cross validation as a function of the number of latent variables for each species partial least squares discriminant analysis model using Savitzky-Golay 2nd derivative preprocessed spectra.

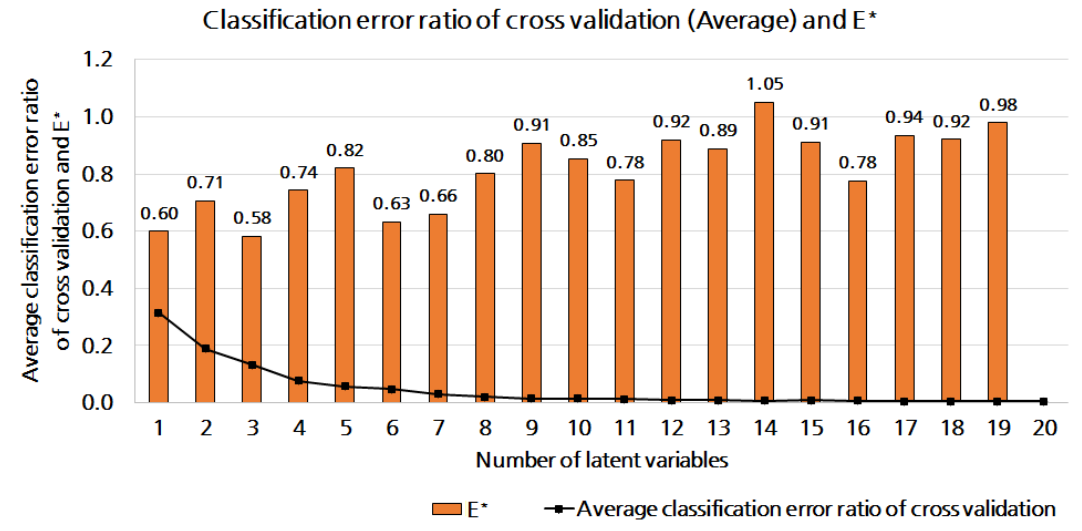


Figure 3-10. Average classification error ratio of cross validation and E* values as a function of the number of latent variables for each species partial least squares discriminant analysis model using Savitzky-Golay 2nd derivative preprocessed spectra.

Fig. 3-11는 Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼을 이용한 각 수종별 부분 최소 자승 판별 분석 모델의 교차 검정 예측치($y_{predicted,\,CV}$)의 분포를 나타내었다. 모든 수종의 부분 최소 자승 판별 모델에 의한 각 집단들의 $y_{predicted,\,CV}$ 는 정규분포로 근사 가능한 것으로 판단되었다.

모든 수종별 판별 분석 모델들이 갖는 해당 수종 집단의 예측치($y_{predicted,\,CV,K}$)는 원 스펙트럼을 이용한 결과와 동일하게 0.5와 1사이의 평균을 가졌으며, 그 외 집단($y_{predicted,\,CV,not\,K}$)의 분포는 0과 -0.5 사이에서 존재하였다. 이는 앞선 두 경우와 차이를 나타내는 부분이었다. 특히 소나무와 잣나무 판별 분석 모델의 경우 원 스펙트럼과 SNV 전처리를 실시한 스펙트럼을 이용한 경우 모두 0.5 미만에서 $y_{predicted,\,CV,K}$ 의 평균을 형성한 것에 비해 Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼을 이용한 경우 0.5와 1사이에서 평균을 형성하여 집단간 분리가 원활해지는 것으로 나타났다.

따라서 Savitkzky-Golay 2nd derivative 전처리는 각 수종별 판별 분석 모델의 판별 정확도를 상당히 개선시키는 효과가 있는 것으로 판단되었다.

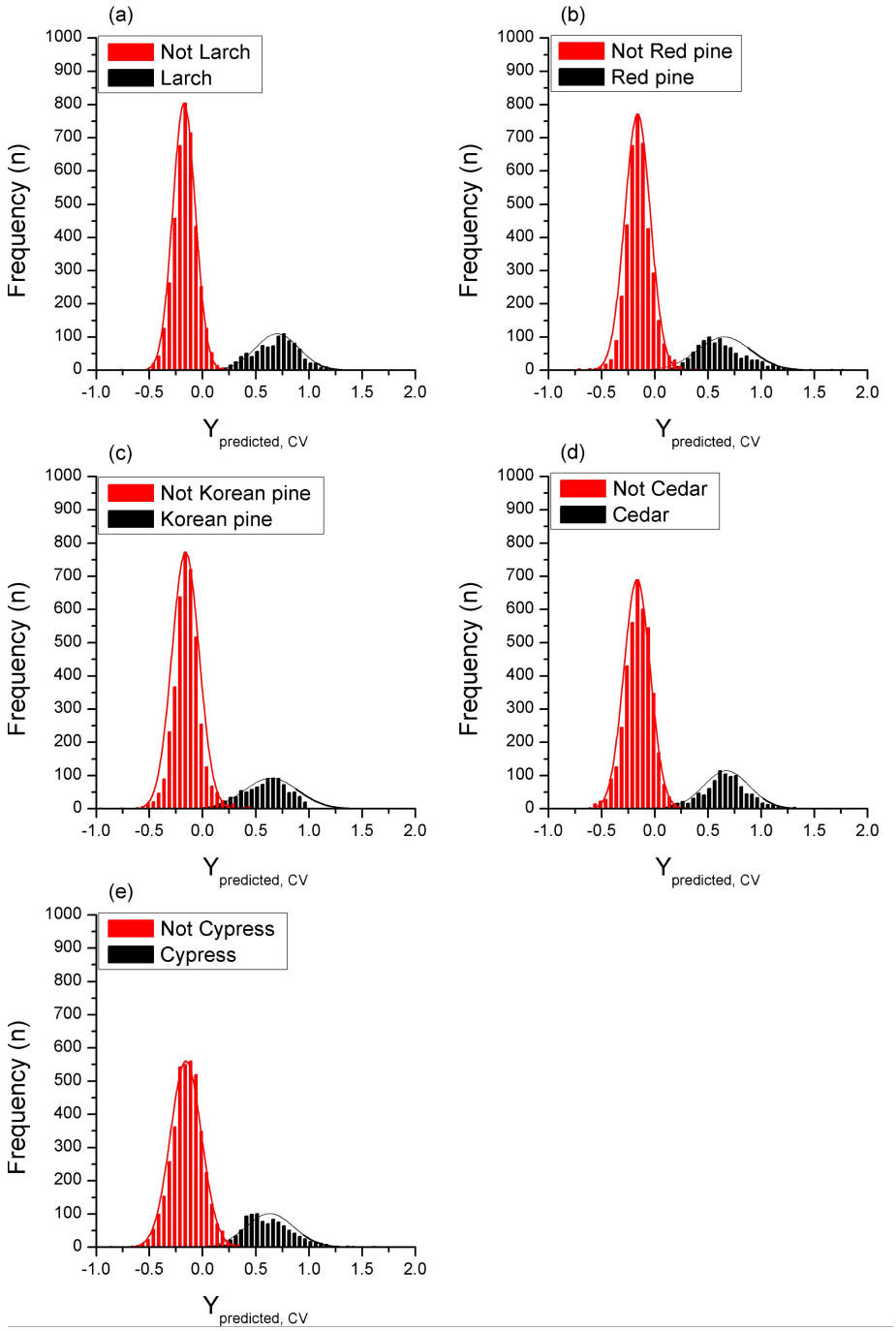


Figure 3-11. Histogram of predicted Y of cross validation by partial least squares discriminant analysis model with 14 latent variables using Savitzky-Golay 2nd derivate preprocessed spectra; (a) model Larch, (b) model Red pine, (c) model Korean pine, (d) model Cedar and (e) model Cypress.

3.2. 원 스펙트럼을 이용한 수종 구분 결과

본 절에서는 수학적 전처리에 따라 3.1.절에서 결정한 요인 수를 갖는 각 수종별 부분 최소 자승 판별 분석 모델들에 스펙트럼을 입력하여 출력된 $y_{predicted,\,CV}$ 값을 이용하여 두 가지 기준에 따라 수종 구분을 실시하였다.

첫 번째는 $y_{predicted,\,CV}$ 값이 0.5 이상인 경우가 1개의 수종 판별 분석 모델에서만 존재하는 경우에만 해당 수종으로 분류하고, 이외의 경우는 기각하는 기준에 따른 수종 구분을 실시하였다. 예를 들어, 관측된 원 스펙트럼 1개를 원 스펙트럼을 이용하여 개발한 각 수종별 판별 분석 모델에 입력하였을 때 낙엽송, 소나무, 잣나무, 삼나무, 편백나무 판별 분석 모델에 의한 $y_{predicted,\,CV}$ 가 각각 (0.75, 0.22, 0.17, -0.04, -0.5)으로 출력되었다면, 다섯 개의 교차 검정 예측치 중 낙엽송 판별모델에 의한 교차 검정 예측치만이 0.75로 나타나 0.5 이상이므로 해당 관측치를 낙엽송으로 판별한다.

두 번째는 $y_{predicted,\,CV}$ 의 분포를 정규분포로 근사하여 구성한 판별 확률 $F_{norm}(y_{predicted,\,CV} \in K)$이 0.5 이상인 경우가 1개의 수종 판별 분석 모델에서만 존재하는 경우에만 해당 수종으로 분류하고, 이외의 경우는 기각하는 기준에 따른 수종 구분을 실시하였다. 예를 들어, 관측된 원 스펙트럼 1개를 원 스펙트럼을 이용하여 개발한 각 수종별 판별 분석 모델에 입력하였을 때 낙엽송, 소나무, 잣나무, 삼나무, 편백나무 판별 분석 모델에 의한 $F_{norm}(y_{predicted,\,CV} \in K)$가 각각 (0.88, 0.20, 0.21, 0.10, 0.09)으로 출력되었다면, 다섯 개의 예측 값 중 낙엽송 판별모델의 확률만이 0.88로 나타나 0.5 이상이므로 해당 관측치를 낙엽송으로 판별한다.

Table 3-2는 원 스펙트럼을 이용하여 개발한 부분 최소 자승 판별 분석 모델에 의해 $y_{predicted,\,CV} \geq 0.5$ 가 1개 모델에서 존재하는 경우만 분류하는 기준에 따른 수종 구분 결과를 나타내었으며, 정확도(Accuracy)는 66.58%로 나타났다. 정밀도(Precision)는 수종에 따라 99.64% ~ 83.78%(낙엽송, 삼나무, 편백, 소나무, 잣나무 순)로 나타났다. 재현율(Recall)은 수종에 따라 84.3% ~ 37.7%(편백, 낙엽송, 삼나무, 소나무, 잣나무 순)로 낮은 수준을 보였다.

Table 3-3은 원 스펙트럼을 이용한 부분 최소 자승 판별 분석 모델에 의해 $F_{norm}(y_{predicted,\,CV} \in K) \geq 0.5$ 가 1개 모델에서 존재하는 경우만 분류하는 기준에 따른 수종 구분 결과를 나타내었으며, 정확도는 67.22%로 나타났다. 정밀도는 수종에 따라 97.81% ~ 81.03%(낙엽송, 삼나무, 편백, 잣나무, 소나무 순)로 나타났다. 재현율은 수종에 따라 86.3% ~ 36.3%(편백, 낙엽송, 삼나무, 잣나무, 소나무 순)로 낮은 수준을 보였다. $y_{predicted,\,CV}$ 의 분포를 고려함으로써 판별 기준치가 조정됨에 따라 미분류는 대폭 감소하였으나 반대로 중복 분류가 대폭 증가하였으며, 이로 인하여 정확도가 미미하게 개선되었다.

원 스펙트럼을 이용한 부분 최소 자승 판별 분석에 의한 수종 구분이 낮은 신뢰도를 나타내는 원인 Fig. 3-5에 나타난 바와 같이 각 수종별 판별 모델에 의한 $y_{predicted,\,CV}$ 의 분포가 중첩되었기 때문으로, 부분 최소 자승 판별 회귀분석에 의해 1 또는 0이 되도록 하는 모조 종속 변수로의 모델화가 원활하지 못하였다는 것을 의미한다. 이는 각 수종별 원 스펙트럼간 유사도가 높았기 때문으로 판단된다.

Table 3-2. Classification results of partial least squares discriminant analysis model with 8 latent variables using raw spectra. (Criteria : $y_{predicted,\,CV} \geq 0.5$)

		Actual Species					Precision (%)
		Larch	Red pine	Korean pine	Cedar	Cypress	
Predicted species	Larch	838	0	3	0	0	99.64
	Red pine	2	481	74	0	0	86.36
	Korean pine	0	71	377	0	2	83.78
	Cedar	1	0	0	790	3	99.50
	Cypress	4	10	3	37	843	93.98
Unassigned		135	381	498	149	146	
Multi-classified	2 species	20	57	44	24	6	
	3 species	0	0	1	0	0	
	4 species	0	0	0	0	0	
	5 species	0	0	0	0	0	
Total		1000	1000	1000	1000	1000	
Recall (%)		83.80	48.10	37.70	79.00	84.30	
Accuracy (%)		66.58					

Table 3-3. Classification results of partial least squares discriminant analysis model with 8 latent variables using raw spectra. (Criteria : $F_{norm}(y_{predicted,\,CV} \in K) \geq 0.5$)

		Actual Species					Precision (%)
		Larch	Red pine	Korean pine	Cedar	Cypress	
Predicted species	Larch	847	0	14	0	5	97.81
	Red pine	0	363	82	3	0	81.03
	Korean pine	3	65	507	1	12	86.22
	Cedar	10	0	0	784	9	97.63
	Cypress	5	5	2	24	860	95.98
Unassigned		41	10	36	8	18	
Multi-classified	2 species	93	549	347	179	82	
	3 species	1	8	12	1	14	
	4 species	0	0	0	0	0	
	5 species	0	0	0	0	0	
Total		1000	1000	1000	1000	1000	
Recall (%)		84.70	36.30	50.70	78.40	86.00	
Accuracy (%)		67.22					

3.3. SNV 전처리가 실시된 스펙트럼을 이용한 수종 구분 결과

Table 3-4는 SNV 전처리가 실시된 스펙트럼을 이용하여 개발한 부분 최소 자승 판별 분석에 모델에 의해 $y_{predicted,\,CV} \geq 0.5$ 가 1개 모델에서 존재하는 경우만 분류하는 기준에 따른 수종 구분 결과를 나타내었으며, 정확도는 62.62%로 원 스펙트럼을 이용한 경우에 비해 오히려 낮았다. 정밀도는 수종에 따라 99.87% ~ 83.28% (삼나무, 낙엽송, 편백, 소나무, 잣나무 순)로 나타났다. 재현율은 수종에 따라 91.1% ~ 27.9% (낙엽송, 삼나무, 편백, 소나무, 잣나무 순)으로 낙엽송 판별을 제외하고는 원 스펙트럼을 이용한 경우에 비해 낮은 수준을 보였다.

Table 3-5는 SNV 전처리가 실시된 스펙트럼을 이용한 부분 최소 자승 판별 분석에 의해 $F_{norm}(y_{predicted,\,CV} \in K) \geq 0.5$ 가 1개 모델에서 존재하는 경우먼 분류하는 기준에 따른 수종 구분 결과를 나타내었으며, 정확도는 63.78%로 $y_{predicted,\,CV} \geq 0.5$ 기준을 이용하는 것에 비해 근소하게 증가하였다. 정밀도는 수종에 따라 98.97% ~ 81.57% (낙엽송, 삼나무, 편백, 잣나무, 소나무순)로 나타났다. 재현율은 수종에 따라 86.8% ~ 30.1% (낙엽송, 삼나무, 편백, 소나무, 잣나무 순)로 낙엽송, 잣나무의 경우를 제외하고는 원 스펙트럼을 이용한 경우에 비해 낮은 수준을 보였다. 따라서 SNV 전처리는 부분 최소 자승 판별 분석에 의한 수종 구분에 부정적인 영향을 나타낸 것으로 판단되었다.

Table 3-4. Classification results of partial least squares discriminant analysis model with 7 latent variables using standard normal variate preprocessed spectra. (Criteria : $y_{predicted,CV} \geq 0.5$)

		Actual Species					Precision (%)
		Larch	Red pine	Korean pine	Cedar	Cypress	
Predicted species	Larch	911	0	9	0	1	98.91
	Red pine	0	476	92	0	1	83.66
	Korean pine	0	54	279	1	1	83.28
	Cedar	0	0	0	749	1	99.87
	Cypress	6	6	1	65	716	90.18
Unassigned		81	424	594	135	273	
Multi-classified	2 species	2	40	25	50	7	
	3 species	0	0	0	0	0	
	4 species	0	0	0	0	0	
	5 species	0	0	0	0	0	
Total		1000	1000	1000	1000	1000	
Recall (%)		91.10	47.60	27.90	74.90	71.60	
Accuracy (%)		62.62					

Table 3-5. Classification results of partial least squares discriminant analysis model with 7 latent variables using standard normal variate preprocessed spectra. (Criteria : $F_{norm}(y_{predicted,CV} \in K) \geq 0.5$)

		Actual Species					Precision (%)
		Larch	Red pine	Korean pine	Cedar	Cypress	
Predicted species	Larch	868	0	8	0	1	98.97
	Red pine	0	301	64	4	0	81.57
	Korean pine	6	45	561	1	22	88.35
	Cedar	2	0	1	650	9	98.19
	Cypress	7	5	3	35	809	94.18
Unassigned		30	3	30	14	33	
Multi-classified	2 species	86	641	323	281	110	
	3 species	1	5	10	15	16	
	4 species	0	0	0	0	0	
	5 species	0	0	0	0	0	
Total		1000	1000	1000	1000	1000	
Recall (%)		86.80	30.10	56.10	65.00	80.90	
Accuracy (%)		63.78					

3.4. Savitzky-Golay 2nd derivative 전처리가 실시된 스펙트럼을 이용한 수종 구분 결과

Table 3-6은 Savitzky-Golay 2nd derivative 전처리가 실시된 스펙트럼을 이용하여 개발한 부분 최소 자승 판별 분석 모델에 의해 $y_{predicted,\,CV} \geq 0.5$ 가 1개 모델에서 존재하는 경우만 분류하는 기준에 따른 수종 구분 결과를 나타내었으며, 정확도는 74.9%로 나타났다. 정밀도는 모든 수종에서 100%로 나타났다. 재현율은 수종에 따라 81.7% ~ 69% (낙엽송, 삼나무, 편백, 소나무, 잣나무 순)로 나타났다.

Table 3-7은 Savitzky-Golay 2nd derivative 전처리가 실시된 스펙트럼을 이용한 부분 최소 자승 판별 분석 모델에 의해 $F_{norm}(y_{predicted,\,CV} \in K) \geq 0.5$ 가 1개 모델에서 존재하는 경우만 분류하는 기준에 따른 수종 구분 결과를 나타내었으며, 이 때의 정확도는 95.18%로 다른 모든 경우와 비교하였을 때 가장 높게 나타났다. 정밀도는 수종에 따라 100% ~ 99.19%로 나타났다. 재현율은 수종에 따라 98.3% ~ 91.5%(낙엽송, 편백, 삼나무, 잣나무, 소나무 순)로 나타났다. 재현율 또한 다른 모든 경우에 비해 가장 양호한 성능을 나타내어 Savitzky-Golay 2nd derivative를 실시한 스펙트럼을 이용하였을 때 가장 신뢰도 높은 수종 판별 분석 모델이 개발되었다. 따라서 Savitzky-Golay 2nd derivative 실시하고, 판별 확률을 이용한 수종 구분을 시행하였을 때 가장 정확한 수준의 수종 구분이 가능한 것으로 판단되었다.

Table 3-6. Classification results of partial least squares discriminant analysis model with 14 latent variables using Savitzky-Golay 2nd derivative preprocessed spectra. (Criteria : $y_{predicted,CV} \geq 0.5$)

		Actual Species					Precision (%)
		Larch	Red pine	Korean pine	Cedar	Cypress	
Predicted species	Larch	817	0	0	0	0	100.00
	Red pine	0	706	0	0	0	100.00
	Korean pine	0	0	721	0	0	100.00
	Cedar	0	0	0	811	0	100.00
	Cypress	0	0	0	0	690	100.00
Unassigned		183	293	279	189	310	
Multi-classified	2 species	0	1	0	0	0	
	3 species	0	0	0	0	0	
	4 species	0	0	0	0	0	
	5 species	0	0	0	0	0	
Total		1000	1000	1000	1000	1000	
Recall (%)		81.70	70.60	72.10	81.10	69.00	
Accuracy (%)		74.90					

Table 3-7. Classification results of partial least squares discriminant analysis model with 14 latent variables using Savitzky-Golay 2nd derivative preprocessed spectra. (Criteria : $F_{norm}(y_{predicted,CV} \in K) \geq 0.5$)

		Actual Species					Precision (%)
		Larch	Red pine	Korean pine	Cedar	Cypress	
Predicted species	Larch	983	0	0	0	2	99.80
	Red pine	0	915	0	0	0	100.00
	Korean pine	0	0	920	0	0	100.00
	Cedar	0	0	0	961	1	99.90
	Cypress	0	0	0	8	980	99.19
Unassigned		0	0	9	0	3	
Multi-classified	2 species	17	84	71	31	14	
	3 species	0	1	0	0	0	
	4 species	0	0	0	0	0	
	5 species	0	0	0	0	0	
Total		1000	1000	1000	1000	1000	
Recall (%)		98.30	91.50	92.00	96.10	98.00	
Accuracy (%)		95.18					

3.5. 부분 최소 자승 판별 분석 모델의 주요 영향 인자

부분 최소 자승 판별 분석을 이용한 수종 구분 시, Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼을 이용하여 모델을 개발하였을 때 각 수종별 판별 모델의 교차 검정 예측치 분포가 가장 원활히 분리되었으며, 가장 높은 정확도로 수종 구분이 가능하였다. 이에 따라 Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼을 이용한 각 수종별 판별 모델에서 예측치 분리에 가장 영향한 파장대역을 탐색하기 위해 variable importance in projection(VIP) score를 활용하였다. VIP score는 각각의 독립변수들이 부분 최소 자승 판별 분석 모델에 미치는 영향을 평가하는 방법(Wold et al., 2002; Akarachantachote et al., 2014; Farres et al., 2015)으로 1 이상의 VIP score를 가지는 파장은 모델의 성능에 큰 비중을 차지한다고 볼 수 있다.

Fig. 3-12는 각 수종별 Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼을 이용하여 개발한 판별 모델의 파장별 VIP score를 나타낸 것이다. 각 수종별 판별 모델에서 가장 VIP score가 높았던 순서대로 5개 파장을 찾은 결과, 낙엽송 판별모델의 경우, 1720, 1739, 1870, 1895, 2304 nm, 소나무 판별모델은 1632, 1674, 1720, 1895, 2304 nm, 잣나무 판별모델은 1632, 1720, 1870, 1895, 2304 nm, 삼나무 판별모델은 1375, 1698, 1895, 1912, 2157 nm, 편백 판별모델은 1632, 1674, 1698, 1895, 2347 nm의 파장대 인근의 스펙트럼 값이 가장 모델에 영향이 높았던 것으로 나타났다.

모든 수종에서 1895 nm 인근의 영향이 높게 나타났으며, 낙엽송, 소나무, 잣나무에서 1720 nm 및 2304 nm 인근의 영향이 높게 나타났다, 소나무 잣나무에서는 1632 nm 인근에서의 영향이 공통적으로 높게 나타났다. 또한, 삼나무와 편백에서는 1696nm 인근에서의 영향이 높게 나타났다. Fig. Ⅱ-4와 Table 1-2을 참고할 때, 1895 nm, 2304 nm 는 아직 목재의 주요 흡광 영역으로 밝혀진 것은 아니나, 스펙트

럼에서는 차이가 발생하였다. 해당 영역은 인근의 큰 흡광영역에 의해 중첩되면서 노출되지 않았거나, 다른 영역에 의해 영향받은 것으로 추측된다. Stefke 등(2008)은 1895 nm 주변은 큰 수분 흡광 영역(1887 ~ 2000 nm)에 의해 중첩되어 드러나지는 않으나 중적외선 영역에서 발생하는 hemicellulose의 C=O stretching 1차 배음대(3600 ~ 3330 cm^{-1})에 기인한 2차 배음대에 의해 약한 흡광이 존재할 것으로 추정된다고 보고하였다. 2304 nm 인근은 결합대로 여러 작용기에 의한 흡광이 중첩되어 발생하는 영역으로, 흡광요인에 대한 특정이 어려웠다. 소나무, 잣나무, 삼나무에서 공통적으로 강한 영향을 나타낸 1632 nm 주변은 cellulose의 O-H stretching에 의해 발생하는 흡광영역으로, compression wood 또는 수피에서 주로 발견되는 흡광영역으로 알려져 있다(Fujimoto and Tsuchikawa, 2010). 낙엽송, 잣나무, 소나무에서 공통적으로 나타난 1720 nm 는 lignin과 hemicellulose의 C-H stretching에 의해 발생하는 흡광영역으로 알려져있다. 삼나무와 편백에서 공통적으로 강한 영향을 나타낸 1698 nm 인근은 lignin의 C-H stretching의 1차 배음대에 의한 흡광영역으로 알려져있다. Savitzky-Golay 2^{nd} deriavtive 전처리에 의해 상기와 같은 수종별 목재의 화학적 성분의 차이가 강조되면서 수종 구분이 원활해지는 것으로 판단된다.

상기에 언급한 파장 뿐 아니라 목재 화학적 성분에서 기인한 여러 파장대역에 의해 수종별 스펙트럼의 차이가 발생하는 것을 확인할 수 있었으나, 정확한 해석을 위해서는 더 많은 연구가 필요할 것으로 판단되었다.

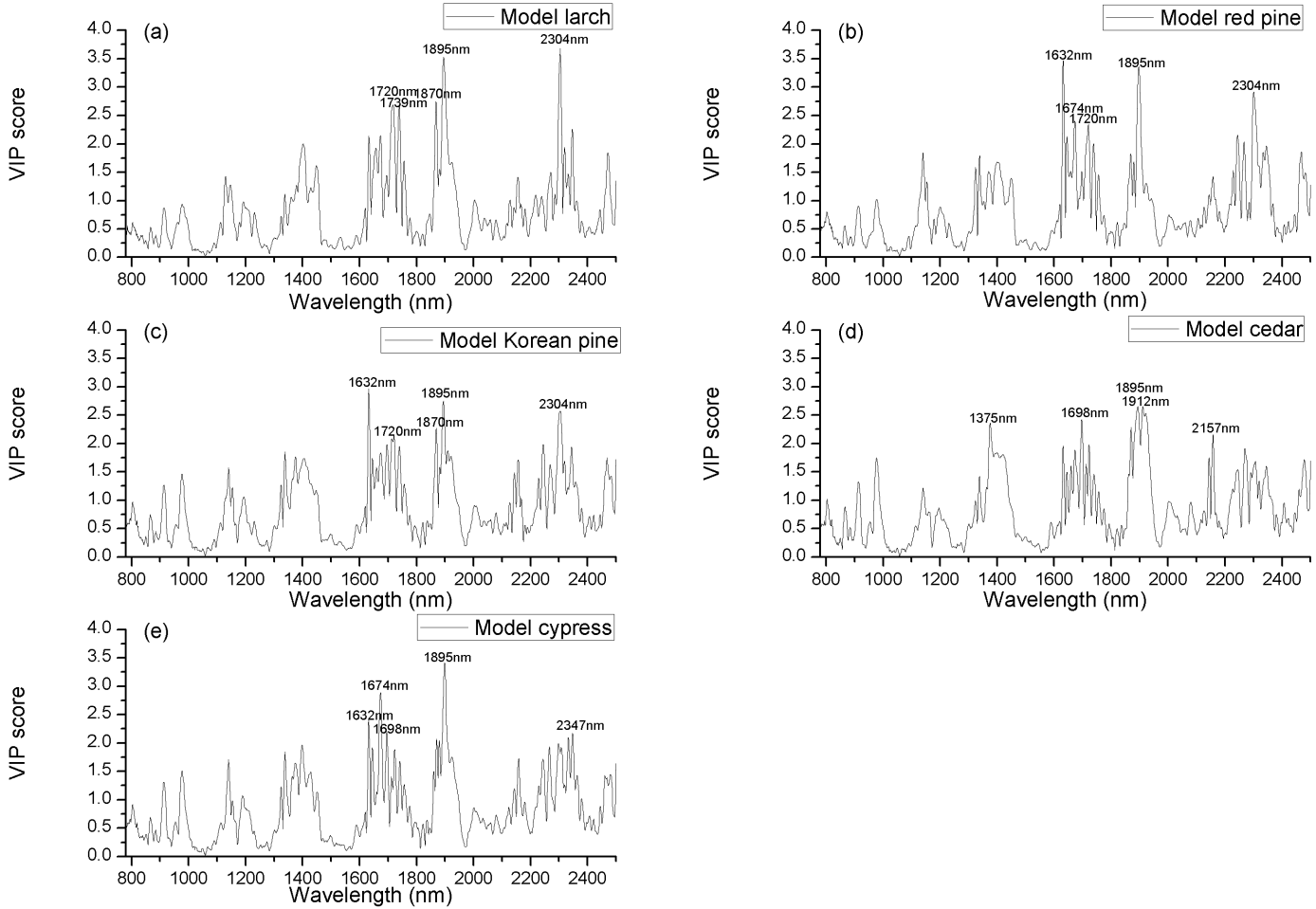


Figure 3-12. Variable importance in projection (VIP) scores for partial least squares discriminant analysis model with 14 latent variables using Savitzky-Golay 2nd derivative preprocessed spectra: (a) model Larch, (b) model Red pine, (c) model Korean pine, (d) model Cedar, (e) model Cypress.

4. 결론

본 장에서는 국산 침엽수 5수종의 제재목 표면에서 측정한 근적외선 흡광도 스펙트럼을 이용하여 원 흡광도 스펙트럼, SNV 전처리를 실시한 흡광도 스펙트럼, Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼 각각을 이용한 부분 최소 자승 판별 모델을 개발하였다. 부분 최소 자승 판별 모델이 포함하는 요인 수에 따른 평균 교차검정 분류 오류율을 이용하여 부분 최소 자승 판별 분석의 최적 모델을 선정하였다.

부분 최소 자승 판별 분석의 교차 검정 예측치($y_{predicted,\,CV}$)를 이용하여 $y_{predicted,\,CV} = 0.5$를 기준으로 판별을 실시하는 경우, 원 스펙트럼을 이용한 판별 모델의 정확도는 66.58%, SNV 전처리를 실시한 스펙트럼을 이용한 판별 모델의 정확도는 62.62%, Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼을 이용한 판별 모델의 정확도는 74.9%로 나타났다.

동일한 결과를 예측치의 분포를 고려하여 확률에 의해 판별을 실시하는 경우, 원 스펙트럼을 이용한 판별 모델의 정확도는 67.22%, SNV 전처리를 실시한 스펙트럼을 이용한 판별 모델의 정확도는 63.78%로 근소한 성능 개선을 나타내었으나, Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼을 이용한 판별 모델의 정확도는 95.18%로 그 성능이 매우 개선되는 것으로 나타났다.

따라서, 제재목에서 측정한 근적외선 스펙트럼을 이용하여 부분 최소 자승 판별 분석을 실시하는 경우, Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼을 이용하여 모델을 개발하되, 교차 검정 예측치의 분포를 반영한 확률적 판별이 수행되는 경우 매우 높은 정확도로 국산 침엽수 5수종의 수종 구분이 가능한 것으로 판단된다.

제 4장

신경망을 이용한 수종 구분

1. 서론

신경망은 입력 데이터로부터 목적에 맞는 특징을 추출하는 가중치를 학습함으로써 예측 또는 분류를 수행하는 알고리즘이다. 목재의 수종 구분을 목적으로 초음파(Jordan et al. 1998), 광학 현미경 사진으로 추출한 해부학적 특징(Esteban et al., 2009; Esteban et al., 2017), 광학 화상(Ravindra et al., 2018), VIS-NIR 스펙트럼(Gong et al., 1997), 근적외선 스펙트럼(Ma et al., 2012; Nisgoski et al., 2017; Lazarescu et al., 2017) 등 다양한 방법으로 측정한 특징들을 이용하여 인공신경망을 적용한 연구들이 수행되었다.

목재의 수종을 구분하기 위해 목재로부터 측정한 근적외선 스펙트럼을 인공신경망에 적용한 연구를 시작한 것은 상대적으로 최근으로, Nisgoski 등(2017)은 활엽수 4수종의 소시편으로부터 측정한 근적외선 스펙트럼의 영역을 구분하여 각각의 파장 영역을 사용한 인공신경망에 의한 수종 구분을 실시한 연구결과를 보고하였다. 그 결과 1000 ~ 2500 nm 영역의 원 스펙트럼을 이용하였을 때, 분류 성능이 가장 정확(100% 정확도)했으며, 스펙트럼 영역을 한정지을 경우, 1820 ~ 2500 nm 대역 또한 높은 분류 정확도(99.5%)를 나타내었다. Lazarescu 등(2017)은 Hem-Fir로 유통되는 제재목 중 Hemlock과 Fir를 구분하기 위해 제재목에서 측정한 근적외선 스펙트럼을 이용하여 인공신경망에 의해 구분하는 연구를 수행하였으며, 두 수종간 구분은 89.8% 이상의 정확도로 가능하다고 보고하였다. 상기의 연구사들을 정리하면, 근적외선 스펙트럼을 이용한 인공신경망에 의한 목재의 수종 구분은 현재 도입 단계에 있으며, 높은 신뢰도로 수종 구분이 가능할 것을 기대할 수 있었다.

이에 본 장에서는 인공신경망과 최근 딥러닝 등으로 각광받고 있는 합성곱 신경망(Convolution neural network)을 1차원 데이터인 근적외선 스펙트럼에 적용한 1차원 합성곱 신경망을 이용한 수종 구분을 실시하고 분류 성능을 평가하였다.

2. 재료 및 방법

2.1. 공시재료

공시재료는 제 1장에서 밝힌 50 × 100 × 600 mm 크기의 낙엽송(Larch, *Larix kaempferi*), 소나무(Red pine, *Pinus densiflora*), 잣나무(Korean pine, *Pinus koraiensis*), 삼나무(Cedar, *Cryptomeria japonica*), 편백(Cypress, *Chamaecyparis obtusa*) 제재목을 수종별로 50개씩 수집하여 기건한 후 활용하였다.

2.2. 근적외선 스펙트럼 측정과 수학적 전처리

공시재료로부터 근적외선 스펙트럼을 획득하는 방법은 제 1장에서 밝힌 방법과 같으며, 동일한 수학적 전처리를 실시하였다.

2.3. 근적외선 스펙트럼의 표준정규화

신경망에 입력되는 스펙트럼 데이터는 각 파장별 스펙트럼 값이 신경망에 미치는 영향이 균등해지도록 파장별로 평균을 0, 표준편차가 1이 되도록 정규화를 실시하였다.

2.4. 인공신경망의 구조와 학습

인공신경망에 의한 수종 구분은 Python 3.6, Keras 2.2.4 및 Tensorflow 1.11.0을 사용하여 실시하였다. 인공신경망에 의한 수종 구분 학습을 실시하기 전, 모델의 신뢰성 검증을 위하여 전체 데이터($N=5000$)를 무작위 선발에 의해 수종별로 8 : 2로 분리하여 4000개 스펙트럼을 학습 세트(Training set)로, 1000개 스펙트럼을 검증 세트(Validation set)로 활용하였다.

인공신경망은 1개의 입력층, 1개의 은닉층, 1개의 출력층으로 구성되었다. 입력층에는 근적외선 스펙트럼의 차원과 같은 1721개 노드, 은닉층에는 64개 노드, 출력층에는 수종의 종류와 같은 5개의 노드를 배치하였다. 이에 따른 인공신경망의 구조와 각 층 사이에 존재하는 인자(가중치 및 bias)의 개수는 Table 4-1과 같다.

Table 4-1. Model structure and parameter of artificial neural network.

Layer type	Input data shape	Number of weights	Number of bias	Total number of parameters
Input layer	1721×1	0	0	0
Hidden layer	64×1	110,144 (1721×64)	64	110,208
Output layer	5×1	320 64×5	5	325

따라서 본 연구에서 설계한 인공신경망은 총 110,533개의 인자를 갖고 있으며, 상기 인공신경망을 이용한 학습이 실시되었다.

은닉층의 활성화 함수로 비선형 활성화 함수인 Rectified linear unit(ReLU)를 사용하였으며, 출력층의 활성화 함수로는 각 수종에 포함될 확률을 출력하기 위하여 Softmax를 사용하였다. 오차 역전파에 의한 학습 최적화 알고리즘으로는 Adam(Kingma and Ba, 2015)이

적용되었으며 이 때의 학습률은 0.001이었다. Adam 등의 확률적 경사 하강법(stochastic gradient descent)에 기반한 최적화 알고리즘들은 학습 세트의 데이터 전체(N_{train})를 순전파시킨 후 오차를 연산하고 이를 역전파하는 것이 아니라, 전체 데이터 중 batch(N_{batch}) 만큼을 순전파하여 오차를 연산하고 이를 역전파하는 과정을 모든 데이터가 실시될 때 까지 반복한다. 따라서 전체 데이터가 순전파와 오차 역전파에 1회씩 사용된 주기를 1 epoch라 할 때, 1 epoch에서도 N_{train}/N_{batch} 횟수만큼 가중치와 bias의 갱신이 시행된다. 이에 본 연구에서는 인공신경망의 빠르고 원활한 학습을 위해 batch size를 32로 설정하였으며, 총 400 epoch 동안 학습을 실시하였다.

인공신경망의 학습 과정 중 학습 세트로의 과적합(Overfitting)을 방지하기 위해 drop-out (Srivastava et al., 2014)이 적용되었다. Drop-out은 Fig. 4-1과 같이 노드의 일부를 제외한 채 가중치를 학습시키는 방법으로, 인공신경망에서 발생할 수 있는 과적합을 방지하면서도 원활한 학습을 실시할 수 있다(Hinton et al., 2012). 이에 본 연구에서는 새로운 epoch 마다 은닉층 노드의 30%를 임의 선택하여 이들과 연결된 가중치를 배제하는 drop-out이 적용된 인공신경망 학습이 실시되었다.

인공신경망의 최적 모델 선정은 전체(400 epch) 학습 과정 중 검증 세트의 오차 지표인 loss가 최소로 나타난 epoch를 선택하였다.

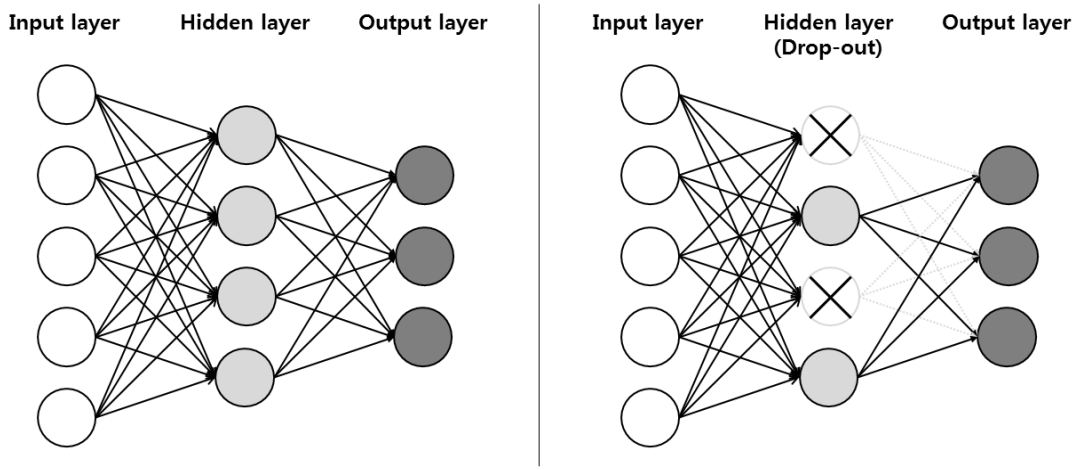


Figure 4-1. Schematic diagram of drop-out in artificial neural networks; left network - an example of artificial neural network with 1 hidden layer.
right network - a thinned artificial neural network by 50% drop-out of the hidden layer on the left neural network.

2.5. 1차원 합성곱 신경망의 구조와 학습

1차원 합성곱 신경망에 의한 수종 구분은 Python 3.6, Keras 2.2.4 및 Tensorflow 1.11.0을 사용하여 실시하였다. 1차원 합성곱 신경망에 의한 수종 구분 학습을 실시하기 전, 모델의 신뢰성 검증을 위하여 전체 데이터를 무작위 선발에 의해 수종별로 8 : 2로 분리하여 8인 4000개 스펙트럼을 학습 세트(N_{train})로, 2인 1000개 스펙트럼(N_{valid})을 검증 세트로 활용하였다.

1차원 합성곱 신경망은 1개의 입력층, 4개의 합성곱 층, 1개의 출력층으로 구성되었다. 입력층에는 근적외선 스펙트럼의 차원과 같은 1721개 노드가 배치되었다. 첫 번째 합성곱 층에서는 3×1 크기의 필터 32채널에 의해 합성곱된다. 두 번째 합성곱 층에서는 9×1 크기의 필터 16채널에 의한 합성곱된다. 세 번째 합성곱 층에서는 27×1 크기의 필터 8채널에 의한 합성곱된다. 네 번째 합성곱 층에서는 1×1 크기의 필터 1채널에 의한 합성곱이 실시된다. 모든 합성곱 연산에서는 차원 유지를 위하여 zero-padding이 적용되었다.

상기와 같은 합성곱 신경망의 구조는 다음을 목표로 설계된 것이다. 첫 번째 합성곱층에서는 1721×1 크기의 입력 스펙트럼을 32종의 작은 크기[3×1]의 필터를 적용함으로써 좁은 영역의 차이를 부각시킨 32개의 1721×1 크기의 스펙트럼으로 분리한다. 두 번째 합성곱층에서는 앞서 분리된 32개의 스펙트럼 마다 각기 다른 중간 크기[9×1]의 필터 16쌍을 이용하여 중간 영역의 차이를 부각시킨 16개의 1721×1 크기의 스펙트럼으로 합성한다. 이에 따라 첫 번째 합성곱층에서 분리되었던 32개의 스펙트럼은 16개로 합성된다. 세 번째 합성곱층에서는 앞서 합성된 16개의 스펙트럼 마다 각기 다른 큰 크기[27×1]의 필터 8쌍을 이용하여 넓은 영역의 차이를 부각시키는 8개의 1721×1 크기의 스펙트럼으로 합성한다. 마지막으로 8개의 1721×1 크기의 스펙트럼을 1개의 1721×1 크기의 스펙트럼으로 합성하기 위해 8개의 스펙트럼마다 각기 다른 (1×1) 크기의 필터를 이용하여 모든 스펙트럼

을 1개의 스펙트럼으로 합성한다. 이 과정을 통해 스펙트럼의 수학적 전처리를 대체하여 자체적으로 수종 분류에 적합한 필터를 학습시킨다. 본 연구에서 실시한 1차원 신경망 구조를 도식화한 형태는 Fig. 4-2와 같다.

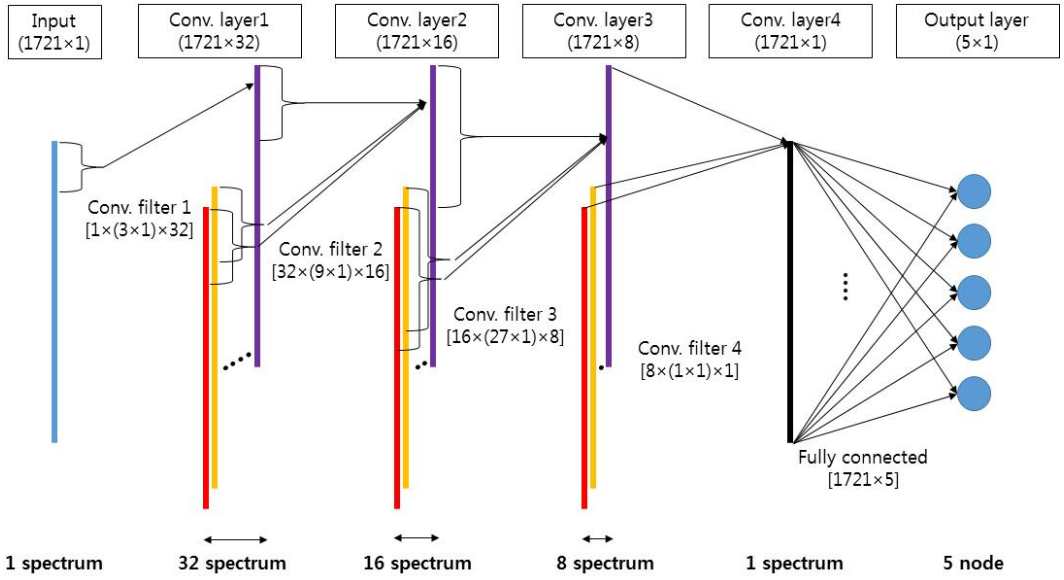


Figure 4-2. Structure of 1 dimensional convolution neural network.

출력층에는 수종의 종류와 같은 5개의 노드를 배치하였다. 이에 따른 1차원 합성곱 신경망의 구조와 각 층 사이에 존재하는 인자(가중치 및 bias)의 개수는 다음과 Table 4-2와 같다.

Table 4-2. Model structure and parameter of 1 dimensional convolution neural network.

Layer type	Input data shape	Number of weights	Number of bias	Total number of parameters
Input layer	1721×1	0	0	0
Convolution layer1	1721×32	96 (1×[3×1]×32)	32	128
Convolution layer2	1721×16	4608 (32×[9×1]×16)	16	4624
Convolution layer3	1721×8	3456 (16×[27×1]×8)	8	3464
Convolution layer4	1721×1	8 (8×[1×1]×1)	1	9
Output layer	5×1	8605 (1721×5)	5	8610

따라서 본 연구에서 실시한 1차원 합성곱 신경망에서는 총 16,835개의 인자를 갖는 신경망에 대한 학습이 실시되었다.

합성곱층의 활성화 함수로 비선형 활성화 함수인 Leaky rectified linear unit(LReLU)를 사용하였으며($\alpha = 0.001$), 출력층의 활성화 함수로 각 수종에 포함될 확률을 출력하기 위하여 Softmax를 사용하였다. 오차 역전파에 의한 학습 최적화 알고리즘으로는 Adam이 적용되었으며 이 때의 학습률은 0.0001이었다. 1차원 합성곱 신경망의 빠르고 원활한 학습을 위해 32개의 batch를 설정하였으며, 총 400 epoch 동안 학습을 실시하였다. 1차원 합성곱 신경망의 학습 과정 중 과적합을 방지하기 위해 drop-out이 적용되었으며, 새로운 epoch 마다 합성곱 신경망 필터의 30%를 임의 선택하여 배제(drop-out)하면서 학습이 실시되었다.

1차원 합성곱 신경망의 최적 모델은 전체(400 epch) 학습 과정 중 검증 세트의 오차 지표인 loss가 최소로 나타난 epoch를 선택하였다.

3. 결과 및 고찰

3.1. 인공신경망을 이용한 수종 구분 결과

3.1.1. 원 스펙트럼을 이용한 인공신경망 모델

원 스펙트럼을 이용하여 시행한 인공신경망의 학습 중 epoch에 따른 정확도와 loss를 Fig. 4-3에 나타내었다. 인공신경망의 학습이 반복적으로 진행됨에 따라 분류 정확도가 지속적으로 증가하다가 수렴하는 경향을 나타냈으며, loss는 지속적으로 감소하다가 수렴하는 경향을 나타내었다. 학습 세트의 정확도가 검증 세트의 정확도에 비해 낮은 것은 은닉층의 노드 중 30%를 배제하고 학습을 실시하는 drop-out이 적용되었기 때문이다. Drop-out에 의하여 학습 세트는 은닉층의 노드 중 70%만을 사용하여 분류 정확도가 평가되고, 검증 세트는 1회 학습이 끝난 후 은닉층의 모든 노드를 이용하여 평가됨에 따라 학습 세트의 정확도가 낮게 평가된다. 원 스펙트럼을 이용한 인공신경망의 학습을 위해 시행된 400 epoch의 학습 중 가장 낮은 loss를 나타낸 지점은 347 epoch로 나타났다. 이 때 검증 세트의 loss는 1.042×10^{-1}였다. 이에 따라 347번째 학습된 가중치와 bias를 최적 모델로 선정하고 이를 이용하여 검증 세트의 수종 구분 정확도를 평가하였다.

Table 4-3은 원 스펙트럼 검증 세트를 347번 학습된 인공신경망에 입력하여 출력된 확률(y_{valid})이 0.5(50%) 이상인 경우가 1개 수종에서만 존재하는 경우만 수종을 구분을 실시하고 이외에는 미분류 하는 방법으로 수종 구분을 실시한 결과를 나타내었다. 검증 세트 수종 구분 정확도는 96.6%로 평가되었다. 정밀도(precision)는 수종에 따라 99.5% ~ 93.14% (낙엽송, 삼나무, 편백, 소나무, 잣나무 순)로 평가되었다. 재현율(recall)은 수종에 따라 100% ~ 92.5%(낙엽송, 편백, 삼나무, 잣나무, 소나무 순)로 나타났다. 원 스펙트럼을 이용한 결과에서는 소나무와 잣나무간, 삼나무와 편백간의 오분류가 발생한 것으로 나타났다.

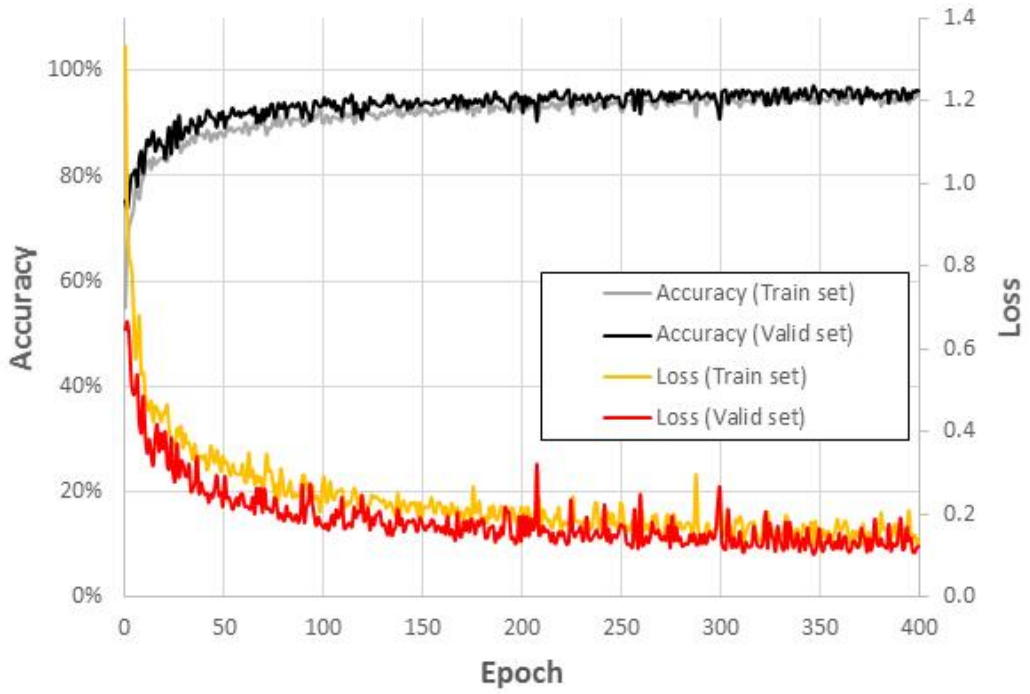


Figure 4-3. Model Accuracy and loss of train/validation set for 400 epochs using artificial neural network based on raw spectra.

Table 4-3. Classification results of artificial neural network (validation set) using raw spectra.

		Actual Species					
		Larch	Red pine	Korean pine	Cedar	Cypress	Precision (%)
Predicted species	Larch	200	0	1	0	0	99.50
	Red pine	0	185	7	0	0	96.35
	Korean pine	0	14	190	0	0	93.14
	Cedar	0	0	0	194	2	98.98
	Cypress	0	0	0	6	197	97.04
Unassigned		0	1	2	0	1	
Multi-classified	2 species	0	0	0	0	0	
	3 species	0	0	0	0	0	
	4 species	0	0	0	0	0	
	5 species	0	0	0	0	0	
Total		200	200	200	200	200	
Recall (%)		100.00	92.50	95.00	97.00	98.50	
Accuracy (%)		96.60					

3.1.2. SNV 전처리를 실시한 스펙트럼을 이용한 인공신경망 모델

SNV 전처리를 실시한 스펙트럼을 이용하여 시행한 인공신경망의 학습 중 epoch에 따른 정확도와 loss를 Fig. 4-4에 나타내었다. 인공신경망의 학습이 반복적으로 진행됨에 따라 분류 정확도가 지속적으로 증가하다가 수렴하는 경향을 나타냈으며, loss는 지속적으로 감소하다가 수렴하는 경향을 나타내었다. 학습 세트의 정확도가 검증 세트의 정확도에 비해 낮은 것은 앞서 설명한 이유와 같다. 인공신경망의 학습을 위해 시행된 400 epoch의 학습 중 가장 낮은 loss를 나타낸 지점은 288 epoch로 나타났다. 이 때 검증 세트의 loss는 2.721×10^{-3}로 원 스펙트럼을 이용한 인공신경망에 비해 크게 감소하였다. 이에 따라 288번 학습된 가중치와 bias를 최적 모델로 선정하고 이를 이용하여 검증 세트의 수종 구분 정확도를 평가하였다.

Table 4-4는 SNV 전처리를 실시한 스펙트럼 검증 세트를 288번 학습된 인공신경망에 입력하여 출력된 확률(y_{valid})이 0.5(50%) 이상인 경우가 1개 수종에서만 존재하는 경우만 수종을 구분을 실시하고 이외에는 미분류 하는 방법으로 수종 구분을 실시한 결과를 나타내었다. 검증 세트 수종 구분 정확도는 99.9%로 평가되었다. 정밀도(precision)는 편백(99.5%)을 제외하고 100%로 평가되었으며, 재현율(recall)은 삼나무를 제외하고 100%로 평가되었다. 원 스펙트럼을 이용한 수종 구분에 비해 매우 높은 정확도, 정밀도 및 재현율을 나타내었으며, 이에 따라 인공신경망을 이용한 수종 구분 시 SNV 전처리가 가지는 성능 개선 효과가 높은 것으로 판단된다.

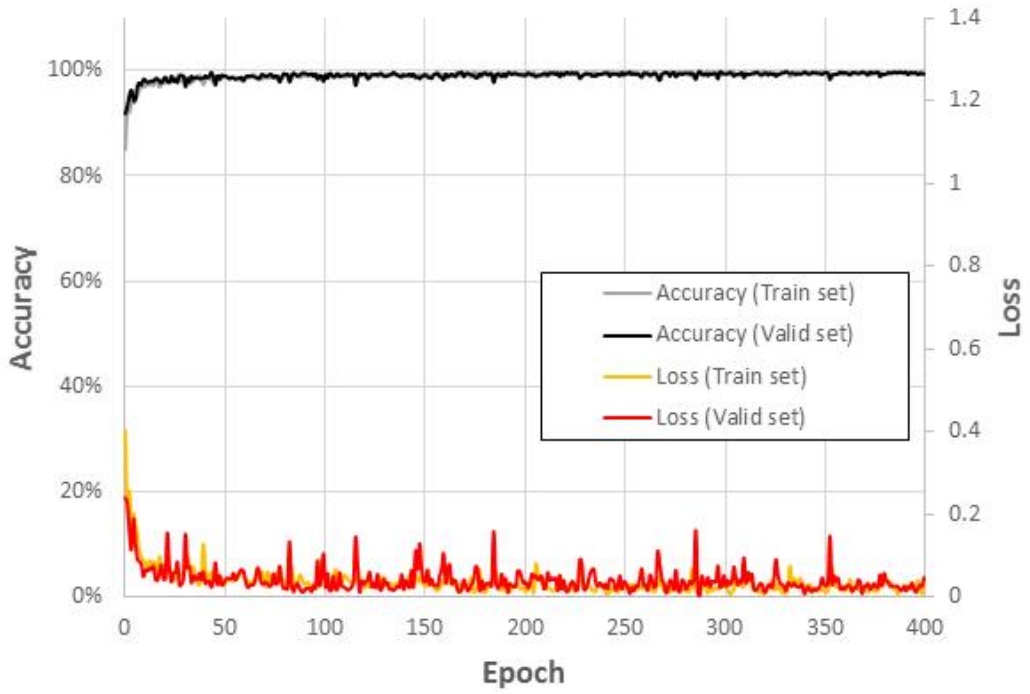


Figure 4-4. Model Accuracy and loss of train/validation set for 400 epochs using artificial neural network based on standard normal variate preprocessed spectra.

Table 4-4. Classification results of artificial neural network (validation set) using standard normal variate preprocessed spectra.

		Actual Species					
		Larch	Red pine	Korean pine	Cedar	Cypress	Precision (%)
Predicted species	Larch	200	0	0	0	0	100.00
	Red pine	0	200	0	0	0	100.00
	Korean pine	0	0	200	0	0	100.00
	Cedar	0	0	0	199	0	100.00
	Cypress	0	0	0	1	200	99.50
Unassigned		0	0	0	0	0	
Multi-classified	2 species	0	0	0	0	0	
	3 species	0	0	0	0	0	
	4 species	0	0	0	0	0	
	5 species	0	0	0	0	0	
Total		200	200	200	200	200	
Recall (%)		100.00	100.00	100.00	99.50	100.00	
Accuracy		99.90					

3.1.3. Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼을 이용한 인공신경망 모델

Savitzky-Golay 2nd derivative 전처리 전처리를 실시한 스펙트럼을 이용하여 시행한 인공신경망의 학습 중 epoch에 따른 정확도와 loss를 Fig. 4-5에 나타내었다. 인공신경망의 학습이 반복적으로 진행됨에 따라 분류 정확도가 단기간에 100%로 수렴하는 경향을 나타냈으며, loss 또한 단기간에 0에 수렴하였다. 인공신경망의 학습을 위해 시행된 400 epoch의 학습 중 가장 낮은 loss를 나타낸 지점은 133 epoch로 나타났다. 이때 검증 세트의 loss는 1.192×10^{-7}로 원 스펙트럼 및 SNV 전처리를 실시한 스펙트럼을 이용한 인공신경망에 비해 크게 감소하였다. 이에 따라 133번 학습된 가중치와 bias를 최적 모델로 선정하고 이를 이용하여 검증 세트의 수종 구분 정확도를 평가하였다.

Table 4-5는 Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼 검증 세트를 133번 학습된 인공신경망에 입력하여 출력된 확률(y_{valid})이 0.5(50%) 이상인 경우가 1개 수종에서만 존재하는 경우만 수종을 구분을 실시하고 이외에는 미분류 하는 방법으로 수종 구분을 실시한 결과를 나타내었다. 검증 세트 수종 구분 정확도, 정밀도, 재현율 모두 100%로 평가되었다. 이에 따라 인공신경망을 이용한 수종 구분 시 Savitzky-Golay 2nd derivative 전처리가 가지는 성능 개선 효과가 높은 것으로 판단된다.

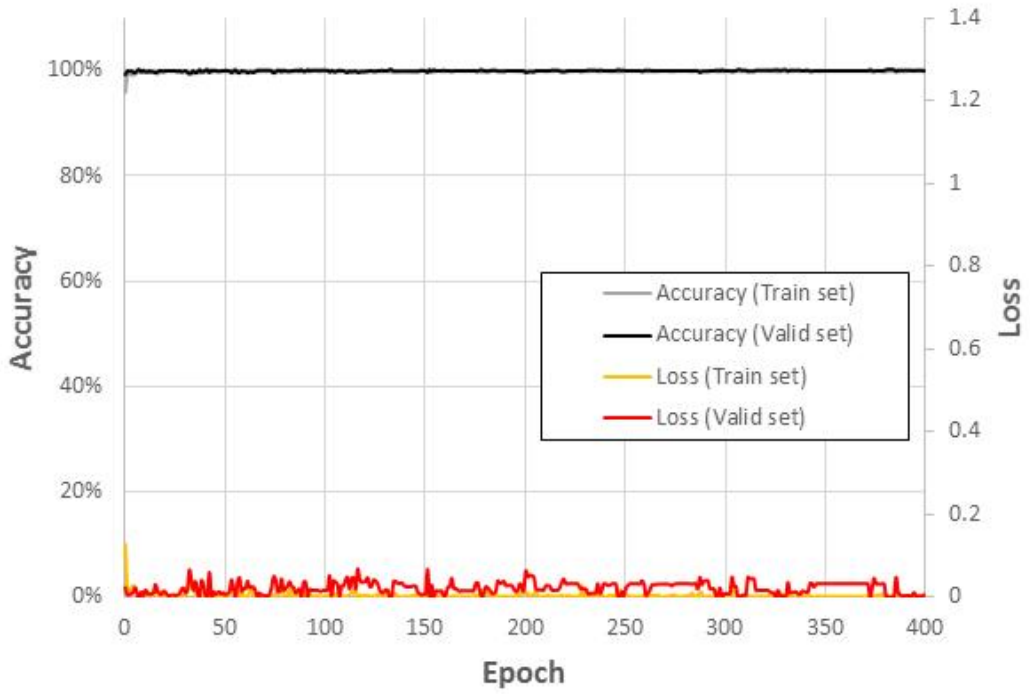


Figure 4-5. Model Accuracy and loss of train/validation set for 400 epochs using artificial neural network based on Savitzky-Golay 2nd derivative preprocessed spectra.

Table 4-5. Classification results of artificial neural network (validation set) using Savitzky-Golay 2nd derivative preprocessed spectra.

		Actual Species					
		Larch	Red pine	Korean pine	Cedar	Cypress	Precision (%)
Predicted species	Larch	200	0	0	0	0	100.00
	Red pine	0	200	0	0	0	100.00
	Korean pine	0	0	200	0	0	100.00
	Cedar	0	0	0	200	0	100.00
	Cypress	0	0	0	0	200	100.00
Unassigned		0	0	0	0	0	
Multi-classified	2 species	0	0	0	0	0	
	3 species	0	0	0	0	0	
	4 species	0	0	0	0	0	
	5 species	0	0	0	0	0	
Total		200	200	200	200	200	
Recall (%)		100.00	100.00	100.00	100.00	100.00	
Accuracy (%)		100.00					

3.2. 1차원 합성곱 신경망을 이용한 수종 구분 결과

3.2.1. 원 스펙트럼을 이용한 1차원 합성곱 신경망 모델

원 스펙트럼을 이용하여 시행한 1차원 합성곱 신경망의 학습 중 epoch에 따른 정확도와 loss를 Fig. 4-6에 나타내었다. 1차원 합성곱 신경망의 학습이 반복됨에 따라 초기 단계에서는 분류 성능이 천천히 상승하였으나, 분류 정확도는 인공신경망에 비해 안정적으로 수렴하는 것으로 나타났다. 학습 세트의 정확도가 검증 세트의 정확도에 비해 낮은 것은 필터 중 30%를 배제하고 학습을 실시하는 drop-out이 적용되었기 때문이다.

원 스펙트럼을 이용한 1차원 합성곱 신경망의 학습을 위해 시행된 400 epoch의 학습 중 가장 낮은 loss를 나타낸 지점은 375 epoch로 나타났다. 이 때 검증 세트의 loss는 5.721×10^{-3}였다. 이에 따라 375번째 학습된 가중치와 bias를 최적 모델로 선정하고 이를 이용하여 검증 세트의 수종 구분 정확도를 평가하였다.

Table 4-6은 원 스펙트럼 검증 세트를 최적 1차원 합성곱 신경망에 입력하여 출력된 확률(y_{valid})이 0.5(50%) 이상인 경우가 1개 수종에서만 존재하는 경우만 수종을 구분을 실시하고 이외에는 미분류 하는 방법으로 수종 구분을 실시한 결과를 나타내었다. 검증 세트 수종 구분 정확도는 99.9%로 평가되었다. 정밀도(precision)는 소나무(99.5%)를 제외하고 모두 100%로 평가되었다. 재현율(recall)은 잣나무(99.5%)를 제외하고 모두 100%로 나타났다. 원 스펙트럼을 이용한 결과에서는 소나무와 잣나무간 오분류가 1개 발생하는 것으로 나타났다. 이러한 결과는 본 연구에서 수행한 모든 원 스펙트럼을 이용한 수종 구분 정확도 중 가장 높았다. 이는 1차원 합성곱 신경망이 분류에 적합한 필터를 자체적으로 개발하면서 발생한 효과로 판단된다.

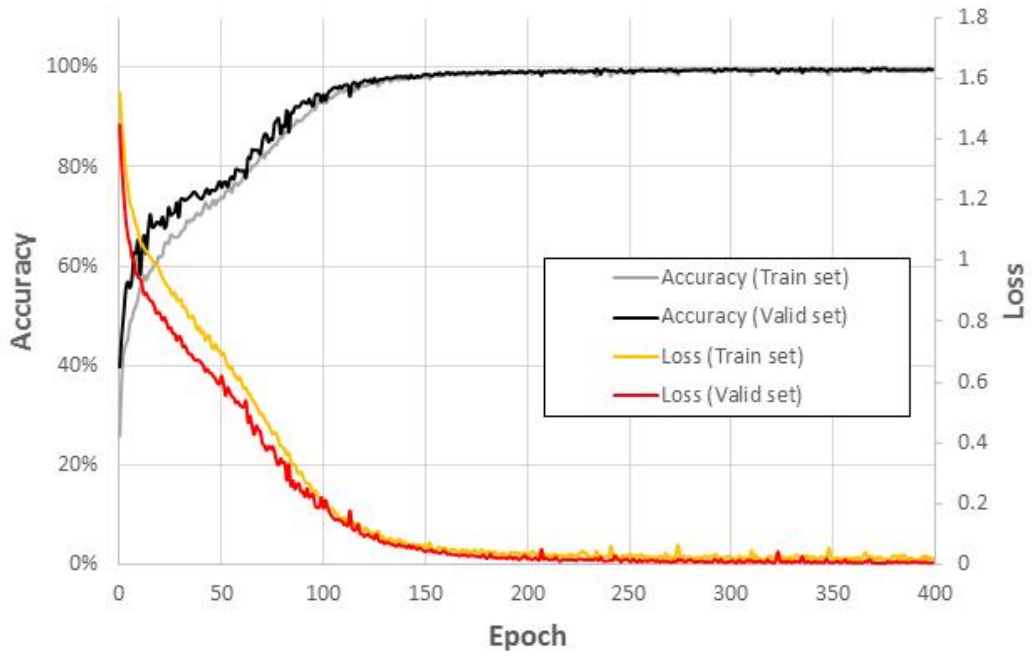


Figure 4-6. Model Accuracy and loss of train/validation set for 400 epochs using 1 dimensional convolution neural network based on raw spectra.

Table 4-6. Classification results of 1 dimensional convolution neural network (validation set) using raw spectra

		Actual Species					
		Larch	Red pine	Korean pine	Cedar	Cypress	Precision (%)
Predicted species	Larch	200	0	0	0	0	100.00
	Red pine	0	200	1	0	0	99.50
	Korean pine	0	0	199	0	0	100.00
	Cedar	0	0	0	200	0	100.00
	Cypress	0	0	0	0	200	100.00
Unassigned		0	0	0	0	0	
Multi-classified	2 species	0	0	0	0	0	
	3 species	0	0	0	0	0	
	4 species	0	0	0	0	0	
	5 species	0	0	0	0	0	
Total		200	200	200	200	200	
Recall (%)		100.00	100.00	99.50	100.00	100.00	
Accuracy (%)		99.90					

3.2.2. SNV 전처리를 실시한 스펙트럼을 이용한 1차원 합성곱 신경망 모델

SNV 전처리를 실시한 스펙트럼을 이용하여 시행한 1차원 합성곱 신경망의 학습 중 epoch에 따른 정확도와 loss를 Fig. 4-7에 나타내었다. 1차원 합성곱 신경망의 학습이 반복됨에 따라 분류 정확도는 원 스펙트럼을 이용하는 모델에 비해 더 빠르게 수렴하는 것으로 나타났다.

SNV 전처리를 실시한 스펙트럼을 이용한 1차원 합성곱 신경망의 학습을 위해 시행된 400 epoch의 학습 중 가장 낮은 loss를 나타낸 지점은 392 epoch로 나타났다. 이 때 검증 세트의 loss는 1.360×10^{-3}로, 원 스펙트럼을 이용하여 학습한 1차원 합성곱 신경망에 비해 약 1/4로 나타났다. 이에 따라 392번째 학습된 가중치와 bias를 최적 모델로 선정하고 이를 이용하여 검증 세트의 수종 구분 정확도를 평가하였다.

Table 4-7은 원 스펙트럼 검증 세트를 최적 1차원 합성곱 신경망에 입력하여 출력된 확률(y_{valid})이 0.5(50%) 이상인 경우가 1개 수종에서만 존재하는 경우만 수종을 구분을 실시하고 이외에는 미분류 하는 방법으로 수종 구분을 실시한 결과를 나타내었다. 검증 세트 수종 구분 정확도는 99.9%, 정밀도는 낙엽송(99.5%)을 제외하고 모두 100%로 평가되었으며, 재현율은 잣나무(99.5%)를 제외하고 모두 100%로 평가되었다. SNV 전처리를 실시한 스펙트럼을 이용한 결과에서는 소나무와 낙엽송간 오분류가 1개 발생하는 것으로 나타났다.

SNV 전처리를 실시한 경우 원 스펙트럼보다 더 빠르게 분류에 적합한 필터를 탐색할 수 있었으며, 검정 세트에 의한 정확도 평가 결과, 원 스펙트럼과 동일하였다.

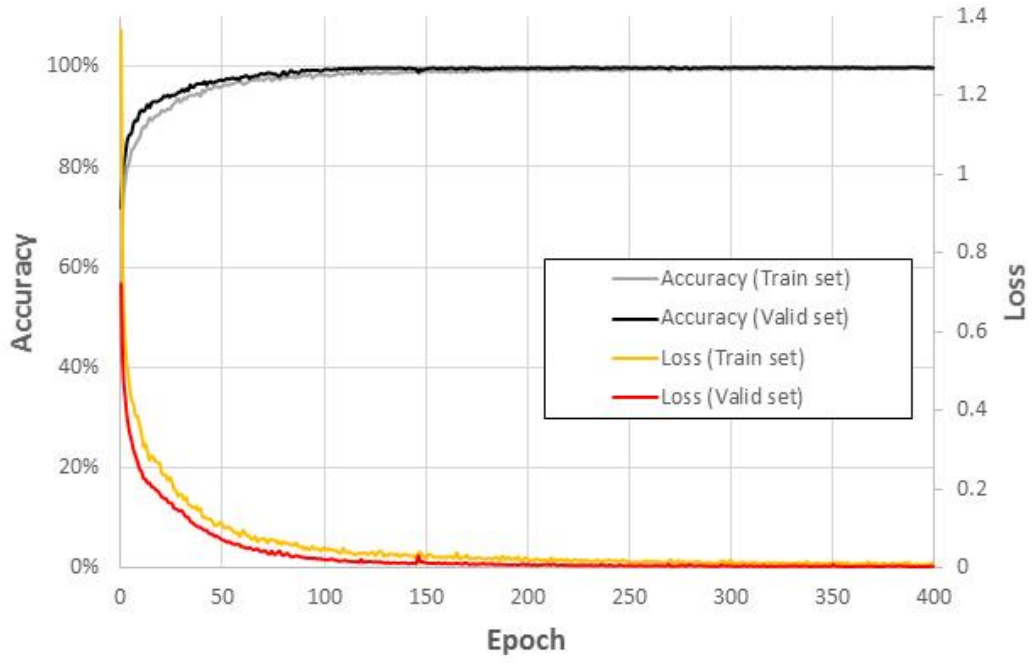


Figure 4-7. Model Accuracy and loss of train/validation set for 400 epochs using 1 dimensional convolution neural network based on standard normal variate preprocessed spectra

Table 4-7. Classification results of 1 dimensional convolution neural network (validation set) using standard normal variate preprocessed spectra.

		Actual Species					
		Larch	Red pine	Korean pine	Cedar	Cypress	Precision (%)
Predicted species	Larch	200	0	1	0	0	100.00
	Red pine	0	200	0	0	0	100.00
	Korean pine	0	0	199	0	0	99.50
	Cedar	0	0	0	200	0	100.00
	Cypress	0	0	0	0	200	100.00
Unassigned		0	0	0	0	0	
Multi-classified	2 species	0	0	0	0	0	
	3 species	0	0	0	0	0	
	4 species	0	0	0	0	0	
	5 species	0	0	0	0	0	
Total		200	200	200	200	200	
Recall (%)		100.00	100.00	99.50	100.00	100.00	
Accuracy (%)		99.90					

3.2.3. Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼을 이용한 1차원 합성곱 신경망 모델

Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼을 이용하여 시행한 1차원 합성곱 신경망의 학습 중 epoch에 따른 정확도와 loss를 Fig. 4-8에 나타내었다. 1차원 합성곱 신경망의 학습이 반복됨에 따라 분류 정확도는 앞선 두 경우보다 더 빠르게 수렴하는 것으로 나타났다. 학습 세트의 정확도가 검증 세트의 정확도에 비해 낮은 것은 은닉층의 노드 중 30%를 배제하고 학습을 실시하는 drop-out이 적용되었기 때문이다.

Savitzky-Golay 2nd derivative 전처리를 실시한 스펙트럼을 이용한 1차원 합성곱 신경망의 학습을 위해 시행된 400 epoch의 학습 중 가장 낮은 loss를 나타낸 지점은 349 epoch로 나타났다. 이 때 검증 세트의 loss는 6.174×10^{-5}로, 원 스펙트럼을 이용한 모델에 비해 약 1/90배 SNV 스펙트럼을 이용한 1차원 합성곱 신경망에 비해 약 1/22배로 나타났다. 이에 따라 394번째 학습된 가중치와 bias를 최적 모델로 선정하고 이를 이용하여 검증 세트의 수종 구분 정확도를 평가하였다.

Table 4-8은 검증 세트를 최적 1차원 합성곱 신경망에 입력하여 출력된 확률(y_{valid})이 0.5(50%) 이상인 경우가 1개 수종에서만 존재하는 경우만 수종을 구분을 실시하고 이외에는 미분류 하는 방법으로 수종 구분을 실시한 결과를 나타내었다. 검증 세트 수종 구분 정확도, 정밀도, 재현율은 모두 100%로 평가되었다. 이에 따라, 수학적 전처리를 미리 실시한 경우 분류에 적합한 필터를 가장 빠르게 학습할 수 있었으며 100% 정확도로 수종 구분이 가능한 것으로 판단되었다.

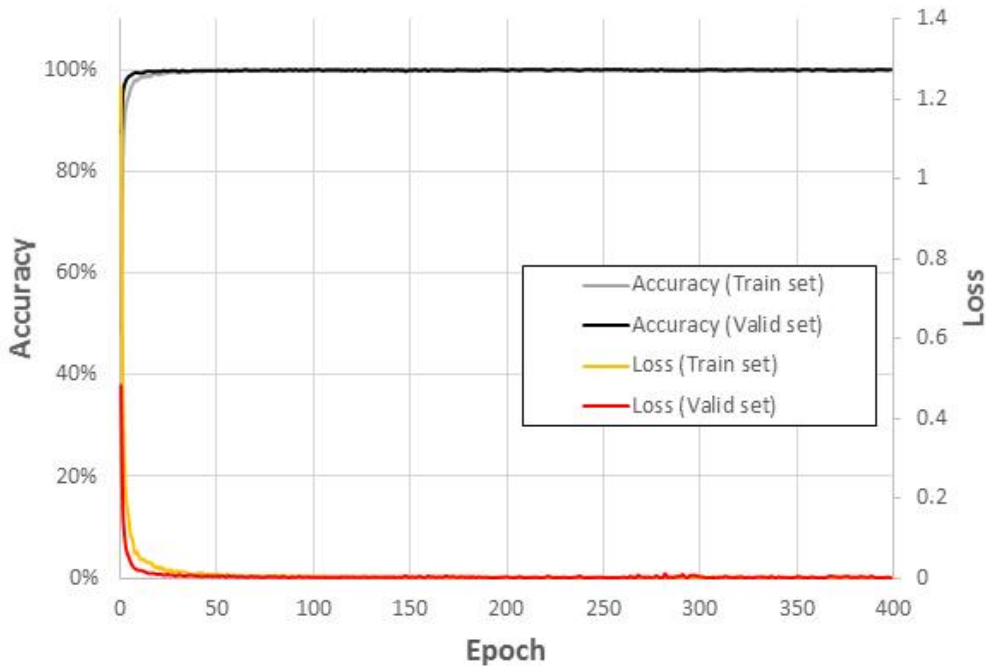


Figure 4-8. Model Accuracy and loss of train/validation set for 400 epochs using 1 dimensional convolution neural network based on Savitzky-Golay 2nd derivative preprocessed spectra.

Table 4-8. Classification results of 1 dimensional convolution neural network (validation set) using Savitzky-Golay 2nd derivative preprocessed spectra.

		Actual Species					
		Larch	Red pine	Korean pine	Cedar	Cypress	Precision (%)
Predicted species	Larch	200	0	0	0	0	100.00
	Red pine	0	200	0	0	0	100.00
	Korean pine	0	0	200	0	0	100.00
	Cedar	0	0	0	200	0	100.00
	Cypress	0	0	0	0	200	100.00
Unassigned		0	0	0	0	0	
Multi-classified	2 species	0	0	0	0	0	
	3 species	0	0	0	0	0	
	4 species	0	0	0	0	0	
	5 species	0	0	0	0	0	
Total		200	200	200	200	200	
Recall (%)		100.00	100.00	100.00	100.00	100.00	
Accuracy (%)		100.00					

4. 결론

본 장에서는 제재목에서 측정된 근적외선 스펙트럼을 이용하여 인공신경망과 1차원 합성곱 신경망을 이용한 수종 구분을 실시하였다. 원 스펙트럼을 이용한 인공신경망 수종 구분의 정확도는 96.6%로 3장의 부분최소 자승 판별 분석의 가장 양호했던 조건에 비해서도 높은 정확도 가지는 것으로 평가되었다. 수학적 전처리를 실시할 시 정확도는 99.5% 이상으로 매우 정확히 수종 구분을 할 수 있는 것으로 판단되었다. 이는 인공신경망 내 비선형 활성화 함수가 가지는 특성 때문으로 판단된다. 즉 집단을 정의하는 모조 종속 변수(dummy dependent variable)와 근적외선 스펙트럼 사이의 비선형적 상관관계에 의해 더 양호한 학습이 이루어지는 것으로 판단되었다.

본 연구에서 실시된 1차원 합성곱 신경망은 학습 과정 중 분류에 적합한 필터를 자체적으로 학습하도록 설계되었으며, 원 스펙트럼과 SNV를 실시한 스펙트럼을 이용한 결과에서 99.9%의 정확도로 수종 구분이 가능하였으며, Savitzky-Golay 2^{nd} derivative 전처리를 실시한 스펙트럼을 이용한 결과 100% 정확도로 수종 구분이 가능하였다. 수학적 전처리를 미리 수행한 스펙트럼을 이용한 경우 더 빠른 속도로 학습이 실시되었다. 이로 비추어 1차원 합성곱 신경망은 근적외선 스펙트럼을 목재의 수종을 구분할 수 있는 형태로 전처리하는 필터를 학습한 것으로 판단된다.

최종적으로, 본 연구에서 적용한 근적외선 스펙트럼을 이용한 여러 수종 구분 방법 중 신경망 이론, 특히 1차원 합성곱 신경망을 활용하는 경우 가장 정확한 수종 구분이 가능한 것으로 판단된다.

제 5장

결론

1. 결론

본 연구에서는 현미경을 이용한 해부학적 세포 비교 분석 또는 DNA 분석 등으로 실시하는 기존의 제재목 수종 식별 방법이 가진 한계(파괴적 시험편 준비 방법, 높은 분석 비용 및 숙련된 전문가 필요)를 극복하기 위하여 제재목의 표면에서 측정한 근적외선 스펙트럼을 이용한 비파괴적 수종 구분을 실시하였다. 공시수종으로는 제재목 생산용으로 공급되는 국산 침엽수 유통량의 대부분을 차지하는 낙엽송, 소나무, 잣나무, 삼나무 및 편백을 선정하였다. 국내 각지의 산림조합에서 생재상태의 공시 수종 제재목을 수집하고, 기건을 실시한 후 제재목의 표면에서 수종 구분을 위한 근적외선 스펙트럼을 획득하였다. 근적외선 스펙트럼을 이용한 제재목의 수종 구분 방법으로 주성분 분석에 의한 군집화 분석, Soft independent modelling of class analogy (SIMCA), 부분 최소 자승 판별 분석, 인공신경망 및 1차원 합성곱 신경망이 활용되었다. 각 수종 구분 방법에 따른 결론은 다음과 같다.

(1) 주성분 분석은 전체 데이터의 분산을 이용하여 데이터들의 방향성인 loading과 각 관측치의 방향성의 크기인 score를 직교 대각화하며 추출하는 분석법이다. 본 연구에서는 주성분 분석에 의해 추출된 전체 스펙트럼 데이터의 방향성 차이인 score의 수종 집단별 군집화를 분석함으로써 수종 구분을 실시하였다.

근적외선 스펙트럼의 3가지 수학적 전처리 조건 (원 스펙트럼, standard normal variate (SNV) 전처리를 실시한 스펙트럼, Savitzky-Golay 2^{nd} derivatives (SG 2^{nd}) 전처리를 실시한 스펙트럼)에 따른 주성분 분석을 실시한 결과, 모든 조건에서 각 수종별 근적외선 스펙트럼의 PC1-PC2 score가 중첩되는 것으로 나타났다. 따라서 주성분 분석에 의한 score의 군집 분석으로는 본 연구에서 활용한 어떠한 수학적 전처리를 실시하더라도 수종 구분이 불가능한 것으로 판단되었다.

(2) SIMCA는 각 집단별 학습 데이터를 주성분 분석하여 발생하는 잔차 분포를 판별 기준으로 삼고, 미지 데이터가 각 집단의 주성분에 의해 근사되면서 발생하는 잔차가 학습 집단의 잔차 분포 내에 포함되는지를 판단하여 분류를 실시한다. 본 연구에서는 근적외선 스펙트럼의 수학적 전처리(원, SNV, SG 2^{nd})에 따라 각 수종 집단별 주성분 분석을 수행하여 SIMCA에 의한 수종 구분을 실시하였다.

원 스펙트럼을 이용한 SIMCA 구분의 정확도는 35.5%, 최소 정밀도는 78.95%, 최소 재현율은 15%로 평가되었으며, 미분류 및 중복 분류가 빈번하게 발생하였다. SNV 전처리를 실시한 스펙트럼을 이용하여 SIMCA 분류를 실시하였을 때의 정확도는 51.9%, 최소 정밀도는 90.67%, 최소 재현율은 19.5%로 나타났으며, 원 스펙트럼을 이용한 결과에 비해 분류 성능이 향상되고 중복 분류의 비율이 현저히 감소하였다. SG 2^{nd} 전처리를 실시한 스펙트럼을 이용한 SIMCA 분류 결과, 정확도는 73%, 최소 정밀도는 98.54%, 재현율은 67.5%로 나타났다(Fig. 5-1).

SIMCA를 이용한 수종 구분은 스펙트럼의 수학적 전처리 조건에 따라 구분 신뢰도의 차이가 나타났으며 SG 2^{nd} 전처리, SNV 전처리, 원 스펙트럼 순서대로 높은 정확도로 수종 구분이 가능하였다. 따라서 근적외선 스펙트럼을 이용하여 SIMCA에 의해 제재목의 수종 구분을 실시하기 위해서는 SG 2^{nd} 전처리를 적용하여 개발한 모델을 활용하는 것이 가장 적합할 것으로 판단되었다.

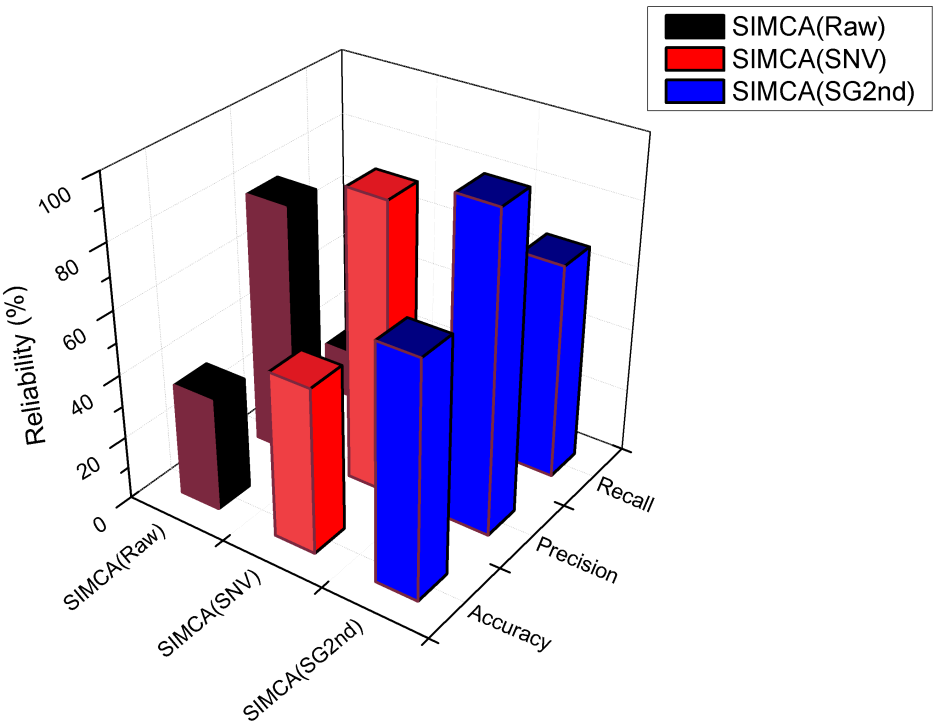


Figure 5-1. Reliability of SIMCA classification model as a function of mathematical preprocessing.

(3) 부분 최소 자승 판별 분석은 구분을 원하는 데이터를 1, 이외의 데이터는 0이 되는 종속변수를 갖도록 하는 부분 최소 자승 모델을 개발하고, 모델별로 출력되는 예측치를 기반으로 분류를 수행하는 방법이다. 본 연구에서는 근적외선 스펙트럼의 수학적 전처리(원, SNV, SG 2^{nd})에 따라 최소 자승 판별 분석 모델을 개발하고 교차 검정 예측치를 기반으로 수종 구분을 실시하였다.

부분 최소 자승 판별 분석에 의한 예측값 $y_{predicted,\,CV} \geq 0.5$를 기준으로 수종 판별을 실시하는 경우, 원 스펙트럼을 이용하여 개발한 부분 최소 자승 판별 모델의 구분 정확도는 66.58%, 최소 정밀도는 83.78%, 최소 재현율은 37.7%로 평가되었다. SNV 전처리를 실시한 스펙트럼을 이용하여 개발한 부분 최소 자승 판별 모델의 구분 정확도는 62.62%, 최소 정밀도는 83.28%, 최소 재현율은 27.9%로 평가되어 원 스펙트럼을 이용한 경우에 비해 낮은 구분 신뢰도를 나타내었다. 따라서 부분 최소 자승 판별 분석 모델 개발 시, SNV 전처리는 판별 신뢰도에 부정적인 효과를 유발하는 것으로 판단되었다. SG 2^{nd} 전처리를 실시한 스펙트럼을 이용하여 개발한 부분 최소 자승 판별 모델의 구분 정확도는 74.9%, 최소 정밀도는 100%, 최소 재현율은 69%로 평가되어 3가지 수학적 전처리 조건 중 가장 개선된 분류 성능을 나타내었다.

부분 최소 자승 판별 분석에 의한 예측값의 분포를 고려하여 확률 밀도 함수를 구성하고, 모델 예측치를 확률로 변환함으로써 각 수종별 예측 확률 $F_{norm}(y_{predicted,\,CV} \in K) \geq 0.5$를 기준으로 수종 판별을 실시하는 경우, 원 스펙트럼을 이용하여 개발한 부분 최소 자승 판별 모델의 구분 정확도는 67.22%, 최소 정밀도는 81.03%, 최소 재현율은 36.3%로 평가되어 예측치 0.5를 기준으로 판별을 수행하는 것에 비해 정확도가 근소하게 개선되었다. SNV 전처리를 실시한 스펙트럼을 이용하여 개발한 부분 최소 자승 판별 모델의 구분 정확도는 63.78%, 최소 정밀도는 81.57%, 최소 재현율은 30.1%로 평가되어 예측치 0.5를 기준으로 판별을 수행하는 것에 비해 정확도가 근소하게 개선되었으나, 여전히 원 스펙트럼을 이용한 모델에 비해 구분 정확도가 낮았다. SG 2^{nd} 전처리를 실시한 스펙트럼을 이용하여 개발한 부분 최소 자승 판별 모델의 구분 정확도는 95.18%, 최소 정밀도는 99.19%, 최소 재현율은 91.5%로 평가되

어 부분 최소 자승 판별 분석을 이용한 수종 구분 조건 중 가장 높은 신뢰도로 수종 구분이 가능하였다(Fig. 5-2).

부분 최소 자승 판별 분석을 이용한 수종 구분은 스펙트럼의 수학적 전처리 조건에 따라 구분 신뢰도의 차이가 나타났으며, SG 2^{nd} 전처리, 원 스펙트럼, SNV 전처리 순서대로 높은 정확도로 수종 구분이 가능한 것으로 평가되었다. 또한 예측치의 분포를 고려하여 확률에 의한 수종 판별을 실시하는 경우, SG 2^{nd} 전처리를 실시한 스펙트럼을 이용하였을 때 수종 구분 성능이 크게 개선되었다. 따라서 근적외선 스펙트럼을 이용한 부분 최소 자승 판별 분석에 의해 제재목의 수종 구분을 실시하기 위해서는 SG 2^{nd} 전처리를 적용하여 개발한 부분 최소 자승 판별 모델의 예측치를 확률로 변환하여 수종 판별을 수행하는 것이 가장 적합한 것으로 판단되었다.

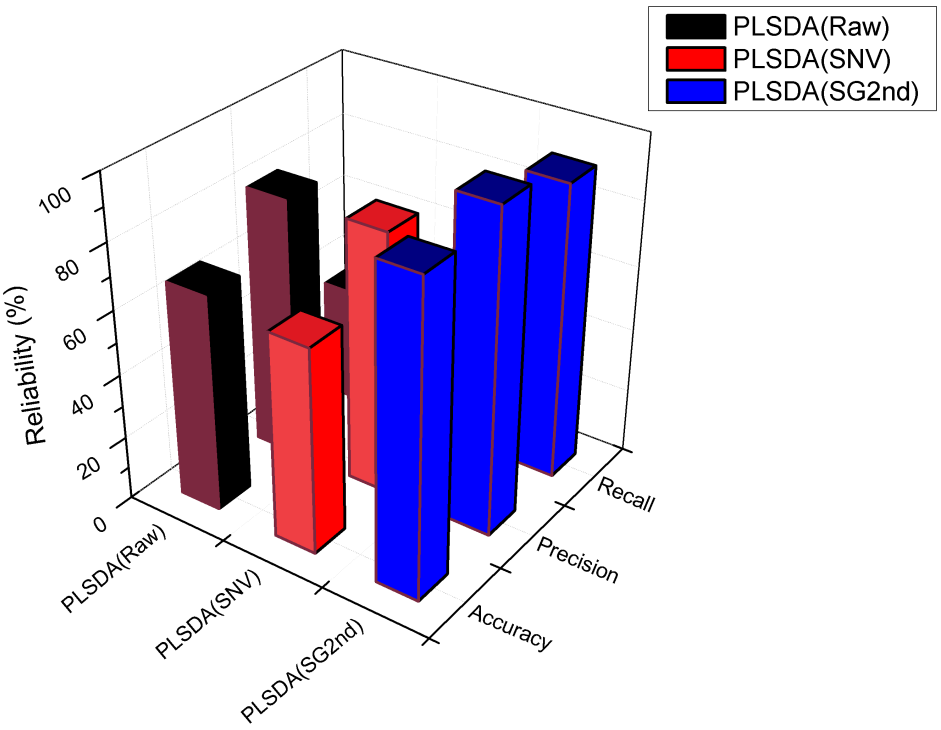


Figure 5-2. Reliability of partial least squares discriminant analysis classification model as a function of mathematical preprocessing (Criteria : $F_{norm}(y_{predicted,\,CV} \in K) \geq 0.5$)

(4) 인공신경망은 입력 데이터로부터 목적에 맞는 특징을 추출하는 가중치를 학습함으로써 예측 또는 분류를 수행하는 알고리즘이다. 본 연구에서는 입력층, 은닉층, 출력층 3개의 층으로 구성된 인공신경망을 이용하여 근적외선 스펙트럼의 수학적 전처리(원, SNV, SG 2^{nd})에 따라 학습된 인공신경망 모델을 이용하여 수종 구분을 실시하였다.

원 스펙트럼을 이용한 인공신경망 모델의 수종 구분 정확도는 96.6%, 최소 정밀도는 93.14%, 최소 재현율은 92.5%로 평가되었으며, 소나무와 잣나무간의 오분류 및 삼나무와 편백 간의 오분류가 일부 발생하였다. SNV 전처리를 실시한 스펙트럼을 이용한 모델의 수종 구분 정확도는 99.9%, 최소 정밀도는 99.5%, 최소 재현율은 99.5%로 평가되었다. 인공신경망을 이용하는 경우 SNV 전처리가 유효한 성능 개선 효과를 발생시켰다. SG 2^{nd} 전처리를 실시한 스펙트럼을 이용하여 개발한 인공신경망 모델의 수종 구분 정확도, 정밀도, 재현율은 모두 100%로 나타났다. 따라서 SG 2^{nd} 전처리 또한 SNV 전처리와 마찬가지로 유효한 성능 개선 효과를 나타내었다(Fig. 5-3).

인공신경망을 이용한 수종 구분은 스펙트럼의 수학적 전처리 조건에 따라 구분 신뢰도의 차이가 나타났으며, SG 2^{nd} 전처리, SNV 전처리, 원 스펙트럼 순서대로 높은 정확도로 수종 구분이 가능한 것으로 평가되었다. 또한, SIMCA 및 부분 최소 자승 판별 분석에 비해 모든 수학적 전처리 조건에서 높은 신뢰도로 수종 구분이 가능한 것으로 평가되었다.

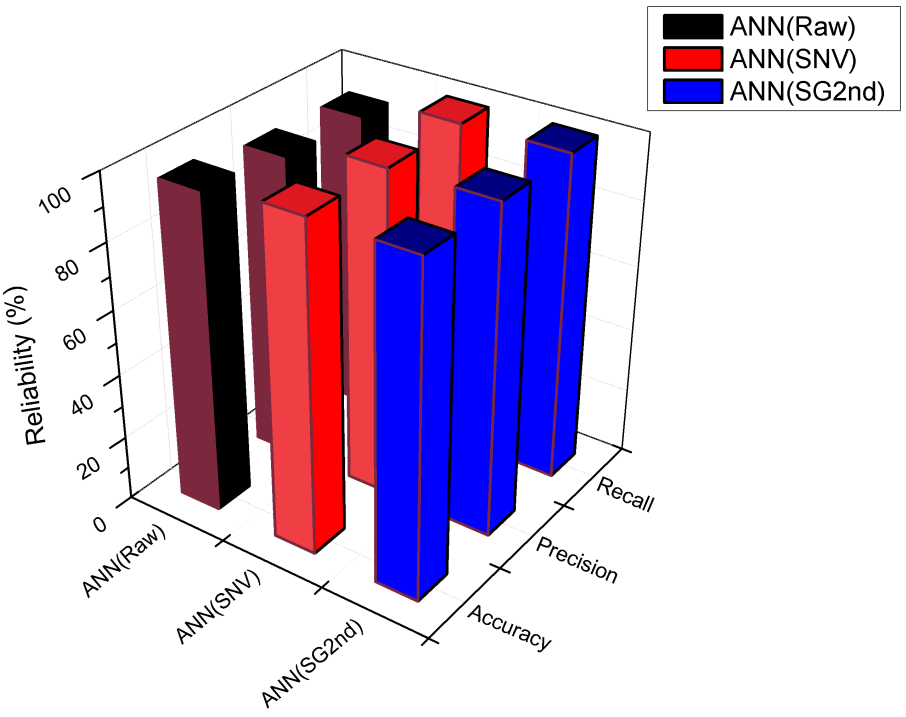


Figure 5-3. Reliability of artificial neural network classification model as a function of mathematical preprocessing.

(5) 1차원 합성곱 신경망은 필터를 이용한 합성곱에 의해 목적에 맞는 특징을 추출하는 필터를 학습함으로써 예측 또는 분류를 수행하는 알고리즘이다. 본 연구에서는 입력층, 4개의 합성곱층, 출력층으로 구성된 1차원 합성곱 신경망을 개발하여 분류에 적합한 4단계의 필터를 합성곱층에서 자체적으로 학습하도록 개발하였다. 근적외선 스펙트럼의 수학적 전처리(원, SNV, SG 2^{nd})에 따라 학습된 1차원 합성곱 신경망 모델을 이용하여 수종 구분을 실시하였다.

원 스펙트럼 및 SNV 전처리를 실시한 스펙트럼을 이용한 1차원 합성곱 신경망 모델의 수종 구분 정확도는 99.9%, 최소 정밀도는 99.9%, 최소 재현율은 99.9%로 평가되었다. SG 2^{nd} 전처리를 실시한 스펙트럼을 이용하여 개발한 인공신경망 모델의 수종 구분 정확도, 정밀도, 재현율은 모두 100%로 나타났다. 1차원 합성곱 신경망을 이용한 수종 구분 시, 근적외선의 수학적 전처리에 관계없이 99.9% 이상의 정확도로 수종 구분을 실시할 수 있는 것으로 평가되었다. 1차원 합성곱 신경망은 근적외선 스펙트럼을 목재의 수종을 구분할 수 있는 형태로 전처리하는 필터를 학습한 것으로 판단된다. 따라서 제재목의 표면에서 획득한 근적외선 스펙트럼을 이용하여 수종 구분을 실시하고자 할 때, 1차원 합성곱 신경망을 이용하다면 최적 수학적 전처리 방법에 대한 탐색 과정 없이도 높은 신뢰도의 모델을 개발할 수 있을 것으로 판단되었다.

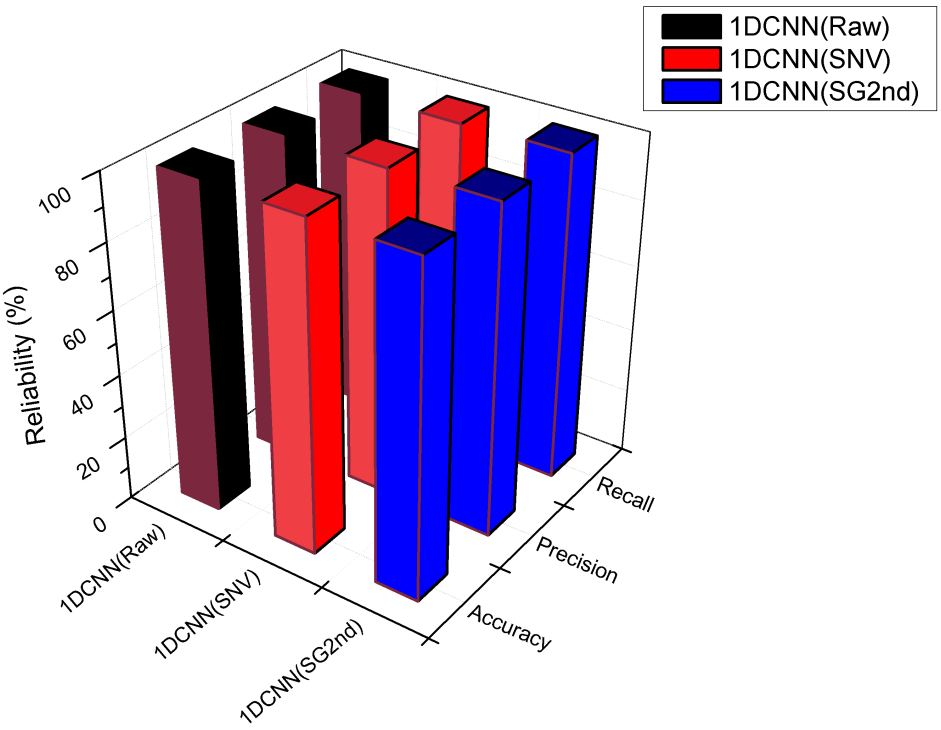


Figure 5-4. Reliability of 1 dimensional convolution neural network classification model as a function of mathematical preprocessing.

2. 향후 과제

본 연구는 사용자가 제재목으로부터 근적외선 스펙트럼을 측정함으로써 간편하고 짧은 시간 내에 수종을 구분하기 위한 방법을 개발하고자 수행하였다. 근적외선 스펙트럼을 이용하여 개발된 수종 구분 방법이 현업에서 활용되기 위하여 필요한 과제를 정리하면 다음과 같다.

(1) 범용성 보완을 위한 시험편 확대

본 연구에서는 국내 각지의 산림조합으로부터 50 × 100 × 600 mm 크기를 갖는 국산 침엽수 5수종(낙엽송, 소나무, 잣나무, 삼나무 및 편백)을 수종별로 50개씩 수집하여 제재목의 표면에서 근적외선 스펙트럼을 측정하였다. 천연 재료인 목재는 생육환경 등에 따라 개체간 변이가 넓어 동일 수종이라 할지라도 물성과 화학적 조성의 차이가 크다. 본 연구에서 수집한 제재목의 수는 한정적이기 때문에 개체간의 변이가 최대한 반영된 근적외선 스펙트럼 데이터를 확보하기 위하여 여러 지역별 산림조합에서 시험편을 수집하였다. 또한 동일한 시험편에 대패가공을 반복적으로 실시함으로써 시험편 내 여러 위치에서 스펙트럼을 측정하였다. 이외에도 모델 개발 시, 수종 구분 모델의 과적합을 방지하기 위한 제약조건을 추가하여 최대한 범용성을 확보하였다.

그럼에도 불구하고, 본 연구에서 수집한 시험편과 근적외선 스펙트럼 데이터는 국내 유통되는 모든 공시 수종 제재목에 대한 대표성을 확보하였음을 보장할 순 없다. 따라서, 본 연구에 의한 결과를 실용적으로 활용하기 위해서는 개체간의 변이를 충분히 포함할만한 수준의 제재목과 근적외선 스펙트럼을 확보할 필요가 있다. 다양한 개체로부터 확보한 근적외선 스펙트럼 데이터를 축적한 후 이를 이용하여 본 연구에서 제시한 방법론에 따라 모델을 개발한다면, 범용성을 확보할 수 있을 것으로 기대된다. 또한, 본 연구에서 선정한 5가지 수종 이외의 수종으로 분류 집단을 확장하기 위해서는 다량의 제재목을 확보하여 개체간의 변이가 충분히 반영된 근적외선 스펙트럼 데이터를 축적함으로써 본 연구의 방법론에 의하여 수종 구분이 가능한 모델의 개발이 가능할 것이다.

(2) 이상치 검사(Outlier detection) 기술

분류 모델은 일반적으로 모델 내 집단임이 확실한 경우에 실시된다. 본 연구에서는 침엽수 5수종(낙엽송, 소나무, 잣나무, 삼나무 및 편백)을 수집하여 이들로부터 측정한 근적외선 스펙트럼을 이용하여 수종 구분 모델을 개발하였기 때문에, 미지 제재목 시험편이 상기 5수종에 포함되는 경우 적용 가능하다. 하지만, 실제 활용 중 본 연구에서 선정한 수종 이외의 제재목의 근적외선 스펙트럼이 모델로 입력될 수 있으며, 극단적으로는 제재목이 아닌 재료에 대한 검사가 실시되는 경우도 발생할 수 있다. 이와 같이 모델 내 집단의 외부에 존재하는 데이터를 이상치(Outlier)라 한다. 이상적으로는 분류 모델이 이상치 데이터를 미분류(Unassigned)로 구분해야 한다. 그러나 실제적으로는 분류 모델이 갖는 판별 특성에 따라 이상치가 특정 수종으로 분류되는 오류가 발생한다.

이에 SIMCA, 부분 최소 자승 판별 분석, 인공신경망, 1차원 합성곱 신경망 수종 구분 모델 이상치에 대한 분류 성능을 확인하였다. 이상치 데이터는 MDF, Ceramic board(단열재의 일종), 북미산 더글라스퍼 제재목으로부터 각각 100개의 근적외선 스펙트럼을 확보하였다. SIMCA, 부분 최소 자승 판별 분석, 인공신경망은 최적 수학적 전처리 조건이었던 SG 2^{nd} 전처리를 실시한 스펙트럼을 이용하여 개발한 모델의 이상치에 대한 분류를 실시하였으며, 1차원 합성곱 신경망은 원 스펙트럼을 이용한 모델을 이용하여 이상치 데이터에 대한 분류를 실시하였다(Table 5-1 ~ Table 5-4).

SIMCA(Table 5-1)는 MDF, ceramic board를 모두 미분류하여 이상치에 대한 미분류가 잘 이루어지는 것을 확인할 수 있었으며, 더글라스퍼도 94%의 데이터가 미분류되어 이상치를 원활하게 검사하였다. SIMCA는 각 집단의 주성분으로의 부합도를 기반으로 잔차를 연산하여 분류를 수행하므로, 이상치에 대한 검사가 분류 알고리즘에 포함되기 때문이다.

반면, 부분 최소 자승 판별 모델(Table 5-2)은 더글라스퍼만 소량(10%) 미분류 하였으며 대부분 중복분류 또는 특정 수종으로 분류가 발생하였다. 이는 부분 최소 자승 판별 분석이 각 수종별 판별 모델(회귀식)에 스펙트럼을 입력하여 출력한 예측치를 기반으로 판별을 수행하기 때문으

로, 이상치 검사에 대한 과정이 부재하기 때문이다. 다만, 부분 최소 자승 판별 분석은 주성분 분석의 발전된 모델이므로 잔차 기반 또는 score 기반 이상치 검정 알고리즘을 추가한다면 이상치 검사가 가능할 것으로 보인다.

인공신경망(Table 5-3) 및 1차원 합성곱 신경망(Table 5-4)에 의한 이상치 검정 결과, 대부분의 이상치 시험편이 특정 수종으로 분류되었다. SG 2^{nd} 전처리를 실시한 스펙트럼을 이용한 인공신경망 모델의 경우 모든 시험편을 잣나무로 분류하여, 이상치에 대한 검사가 불가능하였다.

신경망은 출력층에 위치하는 softmax 함수가 가지는 성질로 인하여 신경망에 의해 각 집단별 예측값(신경망에서는 예측 확률)의 합이 항상 1이(확률의 합은 100%) 된다. 본 연구의 4장에서는 신경망의 예측값을 이용한 수종 판별 기준치를 0.5로 설정하였다(합이 1이 되어야하므로, 0.5를 초과하는 출력값은 항상 1개 이하). 이에 따라 Table 5-3과 Table 5-4는 출력값 0.5를 기준으로 이상치 데이터를 분류한 결과를 나타냈다. 범용적으로 활용 가능한 신경망 모델의 이상치 검정 방법은 연구단계인 것으로 조사되었으나, 현 단계에서는 두 가지 방법에 의해 이상치에 대한 검사가 가능할 것으로 기대된다. 첫 번째는 이외 집단(Others class)을 생성하여 모델 학습을 실시하는 방법이다. 구분을 원하는 5개 수종 집단 이외의 다양한 재료들로부터 가능한 많은 스펙트럼 데이터를 확보하여 이들을 제 6의 집단으로 지정함으로써 모델의 학습을 실시하는 것이다. 이를 통해 이상치 데이터 집단에 대한 분류가 가능할 수 있다. 다만, 이 방법은 '이외 집단'으로 수집할 재료의 범위 및 데이터의 양을 결정할 필요가 있다. 두 번째는 판별 기준치를 상향하여 분류를 수행하는 방법이다. 판별 기준치를 상향하는 것은 더욱 엄격한 기준에 의해 수종을 구분하는 것을 의미한다. 예를 들어, 판별 기준치가 0.5일 때 원 스펙트럼을 이용한 1차원 합성곱 신경망 모델의 수종 구분 결과(Table 4-6)에 비해, 판별 기준치가 0.99일 때의 분류 결과(Table 5-5)는 정확도의 손실은 높지 않았다(99.9%→96.6%). 그러나, 판별 기준치가 높아짐에 따라 이상치 데이터에 대한 검사는 원활해졌다. 모든 신경망 모델이 판별 기준치 상향에 의해 이상치 검사가 가능하지는 않았으나, 하나의 방법으로 활용될 수 있을 것으로 보인다.

Table 5-1. Classification results of outlier set by SIMCA model using SG 2nd preprocessed spectra.

		Actual material		
		MDF	Ceramic board	Douglas-fir
Predicted species	Larch	0	0	0
	Red pine	0	0	5
	Korean pine	0	0	0
	Cedar	0	0	0
	Cypress	0	0	0
Unassigned		100	100	94
Multi-classified	2 species	0	0	1
	3 species	0	0	0
	4 species	0	0	0
	5 species	0	0	0
Total		100	100	100

Table 5-2. Classification results of outlier set by partial least squares discriminant analysis model using SG 2nd preprocessed spectra.

		Actual material		
		MDF	Ceramic board	Douglas-fir
Predicted species	Larch	0	0	43
	Red pine	0	0	0
	Korean pine	0	0	11
	Cedar	0	0	0
	Cypress	0	1	1
Unassigned		0	0	10
Multi-classified	2 species	64	98	28
	3 species	36	1	7
	4 species	0	0	0
	5 species	0	0	0
Total		100	100	100

Table 5-3. Classification results of outlier set by artificial neural network model using SG 2nd preprocessed spectra.

		Actual material		
		MDF	Ceramic board	Douglas-fir
Predicted species	Larch	0	0	0
	Red pine	0	0	0
	Korean pine	100	100	100
	Cedar	0	0	0
	Cypress	0	0	0
Unassigned		0	0	0
Multi-classified	2 species	0	0	0
	3 species	0	0	0
	4 species	0	0	0
	5 species	0	0	0
Total		100	100	100

Table 5-4. Classification results of outlier set by 1 dimensional convolution neural network model using raw spectra.

		Actual material		
		MDF	Ceramic board	Douglas-fir
Predicted species	Larch	100	0	0
	Red pine	0	71	13
	Korean pine	0	0	84
	Cedar	0	0	0
	Cypress	0	29	1
Unassigned		0	0	2
Multi-classified	2 species	0	0	0
	3 species	0	0	0
	4 species	0	0	0
	5 species	0	0	0
Total		100	100	100

Table 5-5. Classification results of 1 dimensional convolution neural network (validation set) model using raw spectra (threshold = 0.99).

		Actual Species					
		Larch	Red pine	Korean pine	Cedar	Cypress	Precision (%)
Predicted species	Larch	200	0	0	0	0	100.00
	Red pine	0	187	0	0	0	100.00
	Korean pine	0	0	189	0	0	100.00
	Cedar	0	0	0	194	0	100.00
	Cypress	0	0	0	0	196	100.00
Unassigned		0	13	11	6	4	
Multi-classified	2 species	0	0	0	0	0	
	3 species	0	0	0	0	0	
	4 species	0	0	0	0	0	
	5 species	0	0	0	0	0	
Total		200	200	200	200	200	
Recall (%)		100.00	93.50	94.50	97.00	98.00	
Accuracy (%)		96.60					

Table 5-6. Classification results of outlier set by 1 dimensional convolution neural network model using raw spectra (threshold = 0.99).

		Actual material		
		MDF	Ceramic board	Douglas-fir
Predicted species	Larch	0	0	5
	Red pine	0	0	0
	Korean pine	0	0	14
	Cedar	0	0	0
	Cypress	0	0	0
Unassigned		100	100	81
Multi-classified	2 species	0	0	0
	3 species	0	0	0
	4 species	0	0	0
	5 species	0	0	0
Total		100	100	100

(3) 함수율

본 연구에서는 생재 제재목을 구입하여 온도 25℃, 상대습도 50 ~ 70%를 유지하는 공간에서 기건한 후, 표면에서 근적외선 스펙트럼을 획득하였다. 따라서 공시 시험편의 함수율은 약 10 ~ 15% 범위로 통제되었다. 근적외선 영역은 수분에 의한 흡광이 강하고 넓게 발생하며, 특히 섬유포화점 이상에서는 매우 큰 흡광이 발생하는 것으로 알려져 있다. 따라서, 본 연구 결과가 다양한 함수율 조건을 갖는 제재목의 수종을 높은 신뢰도로 구분할 것이라 보장하기 어렵다. 다만 근적외선은 목재 내 투과 깊이가 2 mm 이내(Haddadi et al., 2016)로 표면에서 작용하는 것으로 알려져 있고, 제재목의 표면 함수율은 빠르게 기건 상태에 도달할 것으로 예상되어 활용에는 큰 어려움이 없을 것으로 판단된다. 그럼에도 불구하고, 함수율 조건에 영향 받지 않는 수종 구분 모델을 개발할 필요가 있다. 이는 다양한 함수율 조건에서 스펙트럼을 획득하여 모델을 개발하거나, 수분 변화에 독립적인 파장을 탐색하여 해당 파장대역을 이용한 모델을 개발함으로써 극복될 수 있을 것으로 보인다.

(4) 표면 조도에 따른 영향

본 연구에서 활용한 제재목은 대패가공한 후 활용하였으므로, 근적외선 스펙트럼 측정 시 공시 시험편의 표면 조도(surface roughness)는 균일하였다. 근적외선 스펙트럼을 이용한 모델 개발 시에 표면 조도의 영향은 미미하다는 연구 결과들이 보고되어 있으나(Faix and Böttcher, 1992; Hoffmeyer and Pederson, 1995; Schimleck et al., 2005), 제재된 생재의 경우 표면이 매우 거치므로 가급적 표면 조도를 균일화한 후 스펙트럼 측정을 하는 것이 유효할 것으로 판단된다.

(5) 휴대용 장비로의 모델 이식

본 연구에서 사용된 근적외선 분광분석기(SpectraStar 2600XT-R, Unity Scientific, US)는 실험실에서 활용 가능한 설치형 장비다. 본 장비는 680 nm ~ 2600 nm 사이의 스펙트럼을 파장 해상도 1 nm 수준으로 측정할 수 있으며, 신호 대 잡음비(Signal to noise ratio)가 100,000 : 1 이상인 고사양의 분광분석기다.

부피, 무게, 잔적과 같은 제재목의 특성 때문에 제재목의 수종을 구분하기 위하여 제재목을 이동시켜 근적외선 스펙트럼을 측정하기는 어려울 것으로 예상된다. 따라서 현업에서 본 연구에서 개발한 모델을 활용하기 위해서는 휴대용 근적외선 분광분석기에서 적용 가능하도록 수종 구분 모델의 이식이 필요하다. 현재 여러 유통사에서 판매 중인 휴대용(Portable) 근적외선 분광분석기(Table 5-7)는 대개 900 ~ 1700 nm 사이의 파장영역을 6 ~ 10 nm 파장 해상도로 측정할 수 있다. 휴대용 장치들은 본 연구에서 수집한 근적외선 스펙트럼 데이터에 비해 파장 범위와 해상도가 제한되므로, 모델 이식에 따른 수종 구분 성능은 낮아질 것으로 예상된다. 그러나, 수종 구분 모델이 이식된 휴대용 근적외선 분광분석기는 실제 활용성을 높여줄 것으로 기대된다.

Table 5-7. Specifications of several near-infrared spectrometer.

Model	SpectraStar 2600XT-R	NIRvascan	DLP NIRscan Nano EVM	Micro NIR
Producer	Unity Scientific	Allied Scientific Pro	Texas Instruments	VIAVI
Wavelength range (nm)	680 - 2600	900 - 1700	900 - 1700	950 - 1650
Wavelength resolution (nm)	1	10	10	6.2
Signal to noise ratio	100000 : 1	6000 : 1	6000 : 1	25000 : 1

참 고 문 헌

Abdi, H., & Williams, L. J. (2010). Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), 433-459.

Adedipe, O. E., Dawson-Andoh, B., Slahor, J., & Osborn, L. (2008). Classification of Red Oak (Quercus Rubra) and White Oak (Quercus Alba) Wood Using a near Infrared Spectrometer and Soft Independent Modelling of Class Analogies. Journal of Near Infrared Spectroscopy, 16(1): 49-57.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In Selected papers of hirotugu akaike (pp. 199-213). Springer, New York, NY.

Akarachantachote, N., Chadcham, S., & Saithanu, K. (2014). Cutoff threshold of variable importance in projection for variable selection. International Journal of Pure and Applied Mathematics, 94(3): 307-322.

Almeida, M. R., Fidelis, C. H., Barata, L. E., & Poppi, R. J. (2013). Classification of Amazonian rosewood essential oil by Raman spectroscopy and PLS-DA with reliability estimation. Talanta, 117, 305-311.

Alves, A., Santos, A., Rozenberg, P., Paques, L. E., Charpentier, J. P., Schwanninger, M., & Rodrigues, J. (2012). A common near infrared?based partial least squares regression model for the prediction of wood density of Pinus pinaster and Larix× eurolepis. Wood Science and Technology, 46(1-3): 157-175.

Alves, A., Simoes, R., Stackpole, D. J., Vaillancourt, R. E., Potts, B. M., Schwanninger, M., & Rodrigues, J. (2011). Determination of the syringyl/guaiacyl ratio of Eucalyptus globulus wood lignin by near

infrared-based partial least squares regression models using analytical pyrolysis as the reference method. Journal of Near Infrared Spectroscopy, 19(5): 343-348.

Antti, H., Sjostrom, M., & Wallbacks, L. (1996). Multivariate calibration models using NIR spectroscopy on pulp and paper industrial applications. Journal of Chemometrics, 10(5-6): 591-603.

Banerjee, O., Ghaoui, L. E., & d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. Journal of Machine learning research, 9(Mar): 485-516.

Barnes, R. J., Dhanoa, M. S., & Lister, S. J. (1989). Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. Applied spectroscopy, 43(5): 772-777.

Bergo, M. C., Pastore, T. C., Coradin, V. T., Wiedenhoeft, A. C., & Braga, J. W. (2016). NIRS identification of Swietenia macrophylla is robust across specimens from 27 countries. IAWA Journal, 37(3): 420-430.

Bergstrom, B. (2003). Chemical and structural changes during heartwood formation in Pinus sylvestris. Forestry, 76(1): 45-53.

Bishop C. M. (2006) Pattern recognition and machine learning (pp. 78-136). Springer, New York, NY.

Blanco, M., & Villarroya, I. (2002). NIR spectroscopy: a rapid-response analytical tool. TrAC Trends in Analytical Chemistry, 21(4): 240-250.

Braga, J. W. B., Pastore, T. C. M., Coradin, V. T. R., Camargos, J. A. A., & da Silva, A. R. (2011). The use of near Infrared Spectroscopy to Identify solid wood Specimens of Swietenia Macrophylla0 (Cites Appendix II). Iawa Journal, 32(2): 285-296.

da Silva, A. R., Pastore, T. C. M., Braga, J. W. B., Davrieux, F.,

Okino, E. Y. A., Coradin, V. T. R., Camargos, J. A. A. & Do Prado, A. G. S. (2013). Assessment of total phenols and extractives of mahogany wood by near infrared spectroscopy (NIRS).

Dahm, D. J., & Dahm, K. D. (2003). Illustration of Failure of Continuum Models of Diffuse Reflectance. Journal of Near Infrared Spectroscopy, 11(6): 479-485.

Dahm, K. D., & Dahm, D. J. (2004). Relation of Representative Layer Theory to other Theories of Diffuse Reflection. Journal of Near Infrared Spectroscopy, 12(3): 189-198.

De Maesschalck, R., Candolfi, A., Massart, D. L., & Heuerding, S. (1999). Decision criteria for soft independent modelling of class analogy applied to near infrared data. Chemometrics and Intelligent Laboratory Systems, 47(1), 65-77.

Duchi, J., Hazan, E., & Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research, 12(Jul): 2121-2159.

Eom, C.D., Han, Y.J., Chang, Y.S., Park, J.H., Choi, J.W., Choi, I.G., & Yeo, H. (2010). Evaluation of surface moisture content of Liriodendron tulipifera wood in the hygroscopic range using NIR spectroscopy. Journal of the Korean Wood Science and Technology, 38(6): 526-531.

Esteban, L. G., de Palacios, P., Conde, M., Fernandez, F. G., Garcia-Iruela, A., & Gonzalez-Alonso, M. (2017). Application of artificial neural networks as a predictive method to differentiate the wood of Pinus sylvestris L. and Pinus nigra Arn subsp. salzmannii (Dunal) Franco. Wood Science and Technology, 51(5): 1249-1258.

Esteban, L. G., Fernandez, F. G., de Palacios, P. D. P., Romero, R. M., & Cano, N. N. (2009). Artificial neural networks in wood identification: the case of two Juniperus species from the Canary

Islands. IAWA journal, 30(1): 87-94.

Faix, O., & Bottcher, J. H. (1992). The influence of particle size and concentration in transmission and diffuse reflectance spectroscopy of wood. Holz als Roh-und Werkstoff, 50(6): 221-226.

Farres, M., Platikanov, S., Tsakovski, S., & Tauler, R. (2015). Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. Journal of Chemometrics, 29(10): 528-536.

Flæte, P. O., Haartveit, E. Y., & Vadla, K. (2006). Near infrared spectroscopy with multivariate statistical modelling as a tool for differentiation of wood from tree species with similar appearance. New Zealand Journal of Forestry Science, 36(2/3): 382.

Fujimoto, T., & Tsuchikawa, S. (2010). Identification of dead and sound knots by near infrared spectroscopy. Journal of Near Infrared Spectroscopy, 18(6): 473-479.

Fujimoto, T., Kobori, H., & Tsuchikawa, S. (2012). Prediction of wood density independently of moisture conditions using near infrared spectroscopy. Journal of Near Infrared Spectroscopy, 20(3): 353-359.

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. Analytica chimica acta, 185, 1-17.

Gierlinger, N., Jacques, D., Schwanninger, M., Wimmer, R., Hinterstoisser, B., & Paques, L. E. (2003). Rapid prediction of natural durability of larch heartwood using Fourier transform near-infrared spectroscopy. Canadian journal of forest research, 33(9): 1727-1736.

Gong, P., Pu, R., & Yu, B. (1997). Conifer species recognition: An exploratory analysis of in situ hyperspectral data. Remote sensing of Environment, 62(2): 189-200.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep

learning (Vol. 1). Cambridge: MIT press. p.196

Haddadi, A., Hans, G., Leblon, B., Pirouz, Z., Tsuchikawa, S., Nader, J., & Groves, K. (2016). Determination of optical parameters and moisture content of wood with visible‐near infrared spectroscopy. Journal of Near Infrared Spectroscopy, 24(6), 571-585.

He, W., & Hu, H. (2013). Rapid prediction of different wood species extractives and lignin content using near infrared spectroscopy. Journal of Wood Chemistry and Technology, 33(1): 52-64.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. Neural computation, 18(7): 1527-1554.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580.

Hodge, G. R., & Woodbridge, W. C. (2004). Use of near infrared spectroscopy to predict lignin content in tropical and sub-tropical pines. Journal of near infrared spectroscopy, 12(6): 381-390.

Hoffmeyer, P., & Pedersen, J. G. (1995). Evaluation of density and strength of Norway spruce wood by near infrared reflectance spectroscopy. Holz als Roh-und werkstoff, 53(3), 165-170.

Hwang, S. W., Lee, W. H., Horikawa, Y., & Sugiyama, J. (2015). Chemometrics Approach For Species Identification of Pinus densiflora Sieb. et Zucc. and Pinus densiflora for. erecta Uyeki-Species Classification Using Near-Infrared Spectroscopy in combination with Multivariate Analysis. Journal of the Korean Wood Science and Technology, 43(6): 701-713.

Ibraheem, A. K. (2013). An application of artificial neural network classifier for medical diagnosis (Master dissertation, Universiti Tun Hussein Onn Malaysia).

Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial neural networks: A tutorial. Computer, 29(3): 31-44.

Jordan, R., Feeney, F., Nesbitt, N., & Evertsen, J. A. (1998). Classification of wood species by neural network analysis of ultrasonic signals. Ultrasonics, 36, 219-222.

Kienle, A., D'Andrea, C., Foschum, F., Taroni, P., & Pifferi, A. (2008). Light propagation in dry and wet softwood. Optics Express, 16(13): 9895-9906.

Kienle, A., Forster, F. K., & Hibst, R. (2004). Anisotropy of light propagation in biological tissue. Optics Letters, 29(22): 2617-2619.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).

Lazarescu, C., Hart, F., Pirouz, Z., Panagiotidis, K., Mansfield, S. D., Barrett, J. D., & Avramidis, S. (2017). Wood species identification by near-infrared spectroscopy. International Wood Products Journal, 8(1): 32-35.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553): 436-444.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11): 2278-2324.

Li, G., Huang, A., Wang, G., Qin, D., & Jiang, Z. (2007). Rapid determination of Klason lignin content in bamboo by NIR. Guang pu xue yu guang pu fen xi, 27(10): 1977-1980.

Lindgren, F., Geladi, P., & Wold, S. (1993). The kernel algorithm for PLS. Journal of Chemometrics, 7(1), 45-59.

Livni, R., Shalev-Shwartz, S., & Shamir, O. (2014). On the computational efficiency of training neural networks. In Advances in Neural Information Processing Systems (pp. 855-863).

Maas, A. L., Hannun, A. Y., & Ng, A. Y. (2013, June). Rectifier nonlinearities improve neural network acoustic models. In Proc. icml (Vol. 30, No. 1, p. 3).

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5(4): 115-133.

Michell, A. J. (1995). Pulpwood quality estimation by near-infrared spectroscopic measurements on eucalypt woods. Appita Journal, 48(6): 425-428.

Minsky, M., & Papert, S. (1969). Perceptrons. Oxford, England: M.I.T. Press.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th international conference on machine learning (ICML-10) (pp. 807-814).

Nisgoski, S., Carneiro, M. E., & de Muniz, G. I. B. (2015). Influence of sample granulometry on discrimination of Salix species by near infrared. Maderas: Ciencia y Technologia, 17(1): 195-204.

Nisgoski, S., de Oliveira, A. A., & de Muniz, G. I. B. (2017). Artificial neural network and SIMCA classification in some wood discrimination based on near-infrared spectra. Wood Science and Technology, 51(4): 929-942.

Pasquini, C. (2003). Near infrared spectroscopy: fundamentals, practical aspects and analytical applications. Journal of the Brazilian Chemical Society, 14(2): 198-219.

Pastore, T. C. M., Braga, J. W. B., Coradin, V. T. R., Magalhaes, W.

L. E., Okino, E. Y. A., Camargos, J. A. A., Davrieux, F. (2011). Near infrared spectroscopy (NIRS) as a potential tool for monitoring trade of similar woods: Discrimination of true mahogany, cedar, andiroba, and curupixa.

Pearson, K. (1901). Principal components analysis. The London, Edinburgh and Dublin Philosophical Magazine and Journal, 6(2): 566.

Penny, W. D., & Roberts, S. J. (1999). Bayesian neural networks for classification: how useful is the evidence framework?. Neural Networks, 12(6): 877-892.

Perez, N. F., Ferre, J., & Boque, R. (2009). Calculation of the reliability of classification in discriminant partial least-squares binary classification. Chemometrics and Intelligent Laboratory Systems, 95(2): 122-128.

Poke, F. S., Wright, J. K., & Raymond, C. A. (2005). Predicting extractives and lignin contents in Eucalyptus globulus using near infrared reflectance analysis. Journal of Wood Chemistry and Technology, 24(1): 55-67.

Porep, J. U., Kammerer, D. R., & Carle, R. (2015). On-line application of near infrared (NIR) spectroscopy in food production. Trends in Food Science & Technology, 46(2): 211-230.

Qian, N. (1999). On the momentum term in gradient descent learning algorithms. Neural networks, 12(1): 145-151.

Ramalho, F. M., Andrade, J. M., & Hein, P. R. (2018). Rapid discrimination of wood species from native forest and plantations using near infrared spectroscopy. Forest Systems, 27(2), 008.

Ravindran, P., Costa, A., Soares, R., & Wiedenhoeft, A. C. (2018). Classification of CITES-listed and other neotropical Meliaceae wood images using convolutional neural networks. Plant methods, 14(1): 25.

Raymond, C. A., Schimleck, L. R., Muneri, A., & Michell, A. J. (2001). Nondestructive sampling of Eucalyptus globulus and E. nitens for wood properties. III. Predicted pulp yield using near infrared reflectance analysis. Wood Science and Technology, 35(3): 203-215.

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis & Machine Intelligence, (6): 1137-1149.

Richard, M. D., & Lippmann, R. P. (1991). Neural network classifiers estimate Bayesian a posteriori probabilities. Neural computation, 3(4): 461-483.

Robbins, H., & Monro, S. (1951). A Stochastic Approximation Method, Annals Math. Statistics, 22, 400-407.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. nature, 323(6088): 533.

Sandak, A., Sandak, J., & Negri, M. (2011). Relationship between near-infrared (NIR) spectra and the geographical provenance of timber. Wood science and technology, 45(1): 35-48.

Savitzky, A., & Golay, M. J. (1964). Smoothing and differentiation of data by simplified least squares procedures. Analytical chemistry, 36(8): 1627-1639.

SCHIMLECK, L. R., & EVANS, R. (2003). Estimation of air-dry density of increment cores by near infrared spectroscopy. Appita Journal, 56(4): 312-317.

Schimleck, L. R., Michell, A. J., & Vinden, P. (1996). Eucalypt wood classification by NIR spectroscopy and principal components analysis. Appita Journal, 49(5): 319-324.

Schimleck, L. R., Michell, A. J., Raymond, C. A., & Muneri, A. (1999). Estimation of basic density of Eucalyptus globulus using

near-infrared spectroscopy. Canadian Journal of Forest Research, 29(2): 194-201.

Schimleck, L. R., Raymond, C. A., Beadle, C. L., Downes, G. M., Kube, P. D., & French, J. (2000). Applications of NIR spectroscopy to forest research. Appita Journal, 53(6): 458-464.

Schimleck, L. R., Jones, P. D., Clark, A., Daniels, R. F., & Peter, G. F. (2005). Near infrared spectroscopy for the nondestructive estimation of clear wood properties of Pinus taeda L. from the southern United States. Forest Products Journal, 55(12): 21-28.

Siesler, H. W., Ozaki, Y., Kawata, S., & Heise, H. M. (2008). Near-infrared spectroscopy: principles, instruments, applications: John Wiley & Sons.

Soares, L. F., Silva, D. C. D., Bergo, M. C., Coradin, V. T., Braga, J. W., & Pastore, T. (2017). Evaluation of a NIR handheld device and PLS-DA for discrimination of six similar Amazonian wood species. Quimica Nova, 40(4): 418-426.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The Journal of Machine Learning Research, 15(1): 1929-1958.

Stefke, B., Windeisen, E., Schwanninger, M., & Hinterstoisser, B. (2008). Determination of the weight percentage gain and of the acetyl group content of acetylated wood by means of different infrared spectroscopic methods. Analytical chemistry, 80(4), 1272-1279.

Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013, February). On the importance of initialization and momentum in deep learning. In International conference on machine learning (pp. 1139-1147).

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D.,

Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

Thygesen, L. G., & Lundqvist, S.-O. (2000). NIR measurement of moisture content in wood under unstable temperature conditions. Part 1. Thermal effects in near infrared spectra of wood. Journal of Near Infrared Spectroscopy, 8(3): 183-189.

Tsuchikawa, S., & Siesler, H. (2003). Near-infrared spectroscopic monitoring of the diffusion process of deuterium-labeled molecules in wood. Part I: softwood. Applied spectroscopy, 57(6): 667-674.

Tsuchikawa, S., & Tsutsumi, S. (1998). Adsorptive and capillary condensed water in biological material. Journal of materials science letters, 17(8): 661-663.

Tsuchikawa, S., & Yamato, K. (2003). Discriminant Analysis of Wood-Based Materials with Weathering Damage by near Infrared Spectroscopy. Journal of Near Infrared Spectroscopy, 11(5): 391-399.

Tsuchikawa, S., Inoue, K., Noma, J., & Hayashi, K. (2003). Application of near-infrared spectroscopy to wood discrimination. Journal of Wood Science, 49(1): 0029-0035.

Tsuchikawa, S., Torii, M., & Tsutsumi, S. (1998). Directional Characteristics of near Infrared Light in the Process of Radiation and Transmission from Wood. Journal of Near Infrared Spectroscopy, 6(1): 47-53.

Tsuchikawa, S., Yamato, K., & Inoue, K. (2003). Discriminant analysis of wood-based materials using near-infrared spectroscopy. Journal of Wood Science, 49(3): 275-280.

Tyson, J. A., Schimleck, L. R., Aguiar, A. M., Abad, J. I. M., Rezende, G. D., & Otavio Filho, M. (2012). Development of near infrared calibrations for physical and mechanical properties of

eucalypt pulps of mill-line origin. Journal of Near Infrared Spectroscopy, 20(2): 287-294.

Uner, B., Karaman, ?., Tanriverdi, H., & Ozdemir, D. (2009). Prediction of lignin and extractive content of Pinus nigra Arnold. var. Pallasiana tree using near infrared spectroscopy and multivariate calibration. Journal of wood chemistry and technology, 29(1): 24-42.

Via, B. K. (2010). Prediction of oriented strand board wood strand density by near infrared and Fourier transform infrared reflectance spectroscopy. Journal of Near Infrared Spectroscopy, 18(6): 491-498.

Wang, L., Jacques, S. L., & Zheng, L. (1995). MCML-Monte Carlo modeling of light transport in multi-layered tissues. Computer methods and programs in biomedicine, 47(2): 131-146.

Werbos, P. J. (1974). Beyond regression: New tools for prediction and analysis in the behavioral sciences. Doctoral Dissertation, Applied Mathematics, Harvard University, MA.

Whitley, D. (1995). Genetic algorithms and neural networks. Genetic algorithms in engineering and computer science, 3, 203-216.

Williams, C. K., & Barber, D. (1998). Bayesian classification with Gaussian processes. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(12): 1342-1351.

Wold, S. (1976). Pattern recognition by means of disjoint principal components models. Pattern recognition, 8(3): 127-139.

Wold, S., & Sjostrom, M. (1977). SIMCA: A Method for Analyzing Chemical Data in Terms of Similarity and Analogy. In Chemometrics: Theory and Application. pp. 243-282

Wright, J., Birkett, M., & Gambino, M. (1990). Prediction of pulp yield and cellulose content from wood samples using near infrared reflectance spectroscopy. Tappi Journal, 73(8): 164-166.

Wythoff, B. J. (1993). Backpropagation neural networks: a tutorial. Chemometrics and Intelligent Laboratory Systems, 18(2): 115-155.

Yang, S. Y., Park, Y., Chung, H., Kim, H., Park, S. Y., Choi, I. G., Kwon, O., Cho, K. C. & Yeo, H. (2017). Partial least squares analysis on near-infrared absorbance spectra by air-dried specific gravity of major domestic softwood species. Journal of Korean Wood Science and Technology, 45(4): 399-408.

Yang, S. Y., Han, Y., Chang, Y. S., Kim, K. M., Choi, I. G., & Yeo, H. (2013). Moisture Content Prediction Below and Above Fiber Saturation Point by Partial Least Squares Regression Analysis on Near Infrared Absorption Spectra of Korean Pine. Wood and Fiber Science, 45(4): 415-422.

Yang, S.Y., Han, Y., Park, J.H., Chung, H., Eom, C.D., & Yeo, H. (2015). Moisture content prediction model development for major domestic wood species using near infrared spectroscopy. Journal of the Korean Wood Science and Technology, 43(3): 311-319.

Yeh, T.-F., Chang, H.-m., & Kadla, J. F. (2004). Rapid Prediction of Solid Wood Lignin Content Using Transmittance Near-Infrared Spectroscopy. Journal of Agricultural and Food Chemistry, 52(6): 1435-1439.

Yeh, T.-F., Yamada, T., Capanema, E., Chang, H.-M., Chiang, V., & Kadla, J. F. (2005). Rapid screening of wood chemical component variations using transmittance near-infrared spectroscopy. Journal of Agricultural and Food Chemistry, 53(9): 3328-3332.

Zeiler, M. D. (2012). ADADELTA: an adaptive learning rate method. arXiv preprint arXiv:1212.5701.

Zhang, G. P. (2000). Neural networks for classification: a survey. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 30(4): 451-462.

Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A. (2014). Learning deep features for scene recognition using places database. In Advances in neural information processing systems (pp. 487-495).

Abstract

Classification of Wood Species using Near-infrared Spectroscopy and Artificial Neural Networks

Sang Yun Yang
Program in Environmental Material Sciences
Department of Forest Sciences
The Graduate School
Seoul National University

Traditional lumber species identification methods using microscopes identify the species of wood based on anatomical information such as the color, surface features, and anatomic structures. Therefore, highly-trained wood anatomists are essential for anatomical analysis. A DNA analysis is hard to identify the species of lumber due to a difficulty of extracting the nucleus of wood cells, and costs a lot of money for the analysis. In this study, lumber species classification methods based on near-infrared (NIR) spectroscopy and artificial neural networks were developed for the simple and rapid lumber species classification.

Larch (*Larix kaempferi*), red pine (*Pinus densiflora*), Korean pine (*Pinus koraiensis*), cedar (*Cryptomeria japonica*) and cypress (*Chamaecyparis obtusa*) were employed for the study. These five species accounted for the majority of the log supplied to the

domestic lumber production industry. The NIR spectra were acquired from the five species lumber samples, then several algorithms were used for classification.

Principal component analysis of the NIR spectra and soft independent modeling of class analogy (SIMCA) were applied for the species classification. As a result of principal component analysis based on three types of mathematical preprocessing (Raw, standard normal variate (SNV) and Savitzky-Golay 2^{nd} derivatives (SG 2^{nd})), it was impossible to classify the species because PC1-PC2 scores were superposed. The SIMCA model based on SG 2^{nd} preprocessing was determined as the best classification result among SIMCA classification models. The accuracy, minimum precision, and minimum recall of the best model were evaluated as 73.00%, 98.54%, and 67.50%, respectively.

Partial least squares discriminant analysis (PLS-DA) is a multivariate linear regression method. PLS-DA has a dummy dependent variables (1 or 0) depending on the group. There was a difference in the species classification reliability according to the three types of mathematical preprocessing (Raw, SNV, SG 2^{nd}). The PLS-DA model based on SNV preprocessed NIR spectra showed lower classification reliability than that of raw spectra. Thus, the SNV preprocessing affected negatively on the PLS-DA model. The PLS-DA model based on SG 2^{nd} preprocessing was determined as the best PLS-DA classification model. The accuracy, minimum precision, and minimum recall of the best model were evaluated as 74.90%, 100.00%, and 60.00%, respectively.

PLS-DA calculates a classification probability with distributions of the model prediction values when their distributions are considered as

Gaussian probability distribution. PLS-DA models based on raw spectra and SNV preprocessed NIR spectra rarely improved the classification reliability after converting the predicted value to probability. However, PLS-DA model based on SG 2^{nd} preprocessed NIR spectra highly improved (Accuracy = 95.18%, minimum precision = 99.19%, minimum recall = 91.50%). Variable importance in projection analysis was performed to analyze the NIR band that affected the improvement of the PLS-DA model based on SG 2^{nd} preprocessed NIR spectra. 1698 nm which is the light absorbing region of cellulose, 1698 nm which is the light absorbing region of lignin, 1720 nm which is the light absorbing region of lignin and hemicellulose, and 1830 nm and 2304 nm which are not revealed as the light absorbing region by the main component of wood, And it was evaluated as contributing to improvement of species classification performance. There were three distinct peaks including 1632 nm (assigned to the cellulose), 1698 nm (assigned to the lignin), and 1720 nm (assigned to the hemicellulose and lignin) and two not-assigned peaks including 1895 nm and 2304 nm. These peaks were positively affected to the PLS-DA model after SG 2^{nd} preprocessing.

Artificial neural network (ANN) that searches optimum weights for classification from the spectra and 1D convolutional neural network (1D CNN) that searches optimum filters for classification from the input spectra was performed for lumber species classification using the NIR spectra.

ANN has three different layers (Input layer : 1721 nodes, hidden layer : 64 nodes, output layer : 5 nodes) was designed. The classification reliability of ANN models was similar or higher than that of the best classification of PLS-DA based on probabilistic discrimination. Especially, accuracy, precision and recall of ANN

model based on SG 2nd preprocessed NIR spectra was evaluated as 100% each. Also, the classification reliability was improved by mathematical preprocessing in ANN.

In order to reduce dependence on the mathematical preprocessing used in NIR spectroscopy, 1D CNN which finds the optimal mathematical preprocessing. 1D CNN architecture developed in this study has the four 1D convolution layers with different filter sizes and channels. They are arranged to perform preprocessing including spectral filtering, separation, and synthesis. As a result of evaluating classification reliability of 1D CNN model based on raw spectra, the accuracy, minimum precision and minimum recall were 99.90%, 99.50%, and 99.50%, respectively. The reliability of 1D CNN model based on SNV preprocessed spectra was the same as that based on raw spectra. The reliability index was all 100% for 1D CNN based on SG 2nd preprocessed spectra. Finally, it can be concluded that the most reliabile and simple species classification is possible when using neural network theory, especially 1D CNN among the species classification methods using near infrared spectra applied in this study.

Keywords : Near-infared spectroscopy, Species classification, Principal component analysis, Partial least squares discriminant analysis, Artificial neural network, One-dimensional convolutional neural network

Student number : 2013-30337